Molecular and developmental impact of RNA processing on mammalian *Hox* genes

Pedro Miguel Queirós do Patrocínio Patraquim

Submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy at the University of Sussex

April 2015

I hereby declare that this thesis has not been and will not be, submitted in whole or in part to another University for the award of any other degree.

Signed:……..

Pedro Miguel Queirós do Patrocínio Patraquim

UNIVERSITY OF SUSSEX

Pedro Miguel Queirós do Patrocínio Patraquim, DPhil Biology

Molecular and developmental impact of RNA processing on mammalian *Hox* genes

Summary

The *Hox* genes encode a family of evolutionarily conserved transcription factors whose differential expression along head-to-tail triggers distinct programs of cell differentiation along the body axis. Mutations affecting the expression of *Hox* genes disrupt normal development in animals as diverse as insects and mammals. Although the developmental, evolutionary and biomedical relevance of this gene family is indisputable, the understanding of the molecular mechanisms controlling *Hox* gene expression is still incomplete. In particular little is known about the ways *Hox* gene expression is controlled within developmental units such as the insect segments or the rhombomeres in the developing mammalian brain. Previous work in Drosophila showed that different RNA processing events including alternative transcription, alternative splicing and alternative polyadenylation can affect *Hox* gene expression during the development of complex tissues such as the nervous system showing that differential RNA processing contributes to the generation of elaborate *Hox* expression patterns in the fruitfly embryo. Here we explore the impact of RNA processing on the molecular functions and developmental expression of *Hox* genes in mammals. For this we apply a combination of bioinformatic and computational methods complemented by a series of experiments in mammalian cell culture. Our work shows, first, that RNA processing has a pervasive impact on the expression of murine and human *Hox* genes and that specific *Hox* RNA processing reactions are coupled to one another and have evolved in coordination with gene-duplication events. Second, we find that RNA processing affecting several independent *Hox* genes can lead to the generation of Hox protein isoforms that lack a DNA-binding unit (the Homeodomain) suggesting that protein isoforms that are able and unable to bind DNA might be produced during development; furthermore, experiments in cell culture suggest that shorter homeodomain-less isoforms can be generated from longer homeodomain-containing templates suggesting a novel mechanism of RNA processing predicted to substantially impact the biochemical functions of Hox proteins. Third, we find that *Hox* alternative polyadenylation leading to the production of different 3' untranslated regions (3' UTRs) in *Hox* mRNAs can explain the generation of complex spatial patterns of *Hox* expression in the mouse developing limbs and brain. Altogether, our work adds to the current understanding of the molecular control of *Hox* expression during mammalian development, showing that RNA processing can significantly impact the biochemical properties and expression of Hox proteins.

Acknowledgements

First, I would like to thank my PhD supervisor, Claudio Alonso, for all the discussions and endless exchange of ideas, which both fuelled and permeate this thesis. Thank you Joao and Ines for pushing me through the last month in the elaboration of this thesis, and thank you Eva, Ali, Unum, Giulio, Cae, Renanzinho, Juan, Bruce, Maro, Tamara, Mafalda, Sandra, Lambros, Diana John and Dave for the companionship. Thank you Emile and Inaki, for all the smoking breaks and all the scientific discussions than inevitably ensue. Cocas, Vania, Adria and Dani Boss, Gemeas thank you. I hope I can now pay more attention to you. And to the other Dani, Chichas, thank you as well and for the same reasons. Marco, thanks for all the pizza. Thank you Marta for telling me I should do this. Elvira and Maria Pombo, Simone e Zaki, Alex, Patrícia, Marialva, Aninhas Cocas, Zuca, Lima and Sara and your little ones: thanks for being always being there when I come back. Thank you Antoine for that year. Camaradas Igor e Joao Luis: *saudade*.

A special thank you to Aalia Bano for using her talent to solve these biological questions and for being so generous with her data and ideas, and to Sofia Pinho (she knows why) and the rest of the Alonso Lab gang, Stefan, Richard, Wan, Rulo , Agatha. Thank you Martin Ramirez, for being so generous with your limited time so that I could learn basic stuff about cladistics.

Thank you to the FCT for awarding me a PhD fellowship.

Thank you Élio Sucena for always pointing me in the right direction. (over and over). Thank you Luís Carlos Patraquim and Herberto Helder, your words of silk calm us all down. To them and to Jose Rodrigues and Maria Helena, thank you. Finally thank you to my father, Rui, Satanela ,my mother, and to my little brother André. This thesis is dedicated to you.

To Rui Satanela and André, my family

Table of contents

Chapter I **15**

General Introduction **15**

1.1– Preface **16**
1.2- *Hox* genes and the Homeodomain. **17**
1.3- The evolution of mammalian *Hox* clusters. **29**
1.4 - The impact of *Hox* expression on mammalian development. **36**
1.5 - The regulation of mammalian Hox expression at the chromatin and transcriptional levels. **45**
1.6 – RNA-processing in mammalian *Hox* genes. **50**
1.7 – Aims and outcomes of this thesis. **64**

Chapter II **68**

Materials and Methods **68**

2.1 – Batch sequence retrieval from the online database *Ensembl* **69**
2.2 – Estimation of protein divergence rates within Hox paralogue groups. **69**
2.3 – Categorization of mammalian Hox differential RNA processing events for individual protein-coding isoforms **70**
2.4 – Hierarchical-clustering analyses **70**
2.5 – miRNA targeting predictions in the context of alternative Hox 3'UTR formation. **71**
2.6 – Pre-computed protein-domain predictions. **71**
2.7 – Unbiased Hox protein-motif predictions. **72**
2.8 – Computational representation of Hox spatial expression patterns in the developing forelimb of Mus musculus. **73**
2.9 – Bioinformatic search for rhombomere-specific gene-expression in developing hindbrain of Mus musculus. **74**
2.10 – Computational representation of spatial gene expression patterns in the C.elegans germline. **75**
2.11 – Computational search for 3'UTR-enriched motifs. **75**
2.12 – Matching 3'UTR motifs to gene expression patterns using the Subtree pruning and regrafting (SPR) algorithm. **76**
2.13 – Minipreparation of plasmid DNA. **78**
2.14 – Cell culture techniques **78**
2.15 – HEK293-EBNA transfections with plasmid pCMV-Hoxa9. ..... **79**
2.16 – Blocking transcriptional activity in HEK293 cells. **80**
2.17 – RNA extraction. **80**
2.18 – cDNA synthesis. **81**
2.19 – Polymerase Chain Reactions (PCRs) **81**
2.20 – Agarose gel electrophoresis. **83**
2.21 – DNA Sequencing **84**

Chapter III **85**

The production of Hox mRNAs by differential RNA processing in mammals **85**

3.1 – Chapter Overview **86**
3.2 - Results **87**
3.2.1 – Hox genes show evidence of alternative RNA production in mammals. 87
3.2.2 – A catalogue of mammalian alternative *Hox* RNA processing. 96

3.2.3 – The alternative 3'UTRs of mammalian *Hox* mRNAs show a conserved segregation of microRNA (miRNA) target-sites. 105
3.3 – Discussion. 117

Chapter IV 124

The production of Homeodomain-less Hox isoforms by differential RNA processing 124
4.2 – Results 127
4.2.1 – Differential RNA processing produces *Hox* mRNAs that do not encode for the Homeodomain. 127
4.2.2 – M1 and M2 motifs in Hox10 proteins: a case study in the evolution and alternative splicing of functionally important Hox protein motifs....137
4.2.3 – The production of *Hox* mRNAs that do not encode for the Homeodomain is regulated in time and space during *Mus musculus* embryogenesis and adulthood. ..142
4.2.4 – The human gene *Hoxa9* produces mRNAs that lack the Homeodomain. 147
4.2.5 – The *Hoxa9* mRNA sequence is sufficient for the production of alternative mRNAs that lack the Homeodomain. 153
4.2.6 - *Hoxa9* produces mRNAs that lack the Homeodomain in a transcriptional-dependent manner. 157
4.2.7 – All major Transcription Factor families produce mRNA isoforms that do not encode for a DNA-binding domain in a conserved manner across metazoans. 163
4.3 – Discussion. 167

Chapter V 171

The role of Hox 3'UTRs in the coordination of spatial gene expression during mammalian development 171
5.1 – Chapter Overview 172
5.2 – Results 173
5.2.1 – Mammalian *Hox* 3'UTR contain a host of shared, conserved sequence motifs. 173
5.2.2 – Shared *Hox* 3'UTR motifs significantly match mRNA co-expression profiles in the mouse forelimb. 178
5.2.3 –Forelimb-enriched *Hox* 3'UTR motifs include RNA secondary-structures, as well as RBP binding-sites. 187
5.2.4 – 3'UTR motifs of evolutionarily unrelated genes match spatial mRNA expression in the mouse hindbrain. 191
5.2.5 – Validation of the SPR method as a test for 3'UTR-mediated coordination of gene expression: the *Caenorhabditis elegans* germline. 197
5.3 - Discussion 200

Chapter VI 207

***General Discussion* 207**
6.1 – General Discussion 208
6.2 – Paralogous *Hox* genes share patterns of differential RNA processing in mammals. 209
6.3 – Differential RNA processing diversifies Hox protein-sequences in mammals. 213
6.4 – Co-expressed *Hox* mRNAs share a host of sequence motifs in 3' untranslated regions in the developing hindbrain and limb of mammals. 216
6.6 – Concluding remarks 219

References 222

List of Figures

Figure 1.1 Figure 1.1 - *Hox* clusters are conserved across animals and confer positional information across the A-P axis during development_____18

Figure 1.2 The mutation of *Hox* genes causes homeotic transformations in *Drosophila melanogaster* and *Mus musculus*_____20

Figure 1.3 Hox genes contain the Homeobox, which encodes for the Homeodomain_26

Figure 1.4 The evolution of vertebrate *Hox* gene clusters by gene duplication_____32

Figure 1.5 *Hox* genes control the morphogenesis of the axial skeleton, limbs and hindbrain of mammals_____39

Figure 1.6 *HoxD* genes are subjected to chromatin and transcriptional regulation in the developing mammalian limb. _____43

Figure 1.7 *Hox* gene expression is subjected to a host of regulatory levels_____47

Figure 1.8 *Ubx* mRNAs undergo regulated alternative splicing and polyadenylation during *Drosophila melanogaster* development_____57

Figure 3.1 Mammalian *Hox* genes display conserved production of alternative mRNAs by differential RNA processing_____88

Figure 3.2 Accelerated protein evolution in posterior *Hox* PGs is strongly associated with the production of alternative mRNAs_____94

Figure 3.3 Specific modes of differential *Hox* RNA processing show distinct links to transcriptional regulation_____99

Figure 3.4 Differential *Hox* RNA processing involves the coordination of multiple regulatory levels and two distinct modes_____103

Figure 3.5 Experimentally validated miRNAs are predicted to bind to more numerous and stronger targets in distal *Hox* 3'UTRs_____110

Figure 3.6 Alternative cleavage and polyadenylation generates developmental and evolutionary compartments in *Hox* 3'UTRs_____115

Figure 4.1 An unbiased search for Hox protein motifs recovers key Hox domains involved in the molecular function of Hox proteins_____131

Figure 4.2 Hierarchical clustering of Hox protein motifs groups alternative Hox isoforms of the same paralogue group_____133

Figure 4.3 Alternative splicing of *Hoxa10* generates atavistic Hox10 protein isoforms_____139

Figure 4.4 *Mus musculus Hox* genes *Hoxa1*, *Hoxa9*, *Hoxc4*, *Hoxb9* and *Hoxb1* produce mRNAs that do not encode for a Homeodomain in a developmentally regulated manner _____144

Figure 4.5 Alternative splicing of *Hoxa9* produces Homedomain-encoding and Homeodomain-less mRNA isoforms in mammals_____149

Figure 4.6 The Homeodomain-encoding cDNA of Hoxa9 is sufficient to produce the Homeodomain-less mRNA upon overexpression_____154

Figure 4.7 Transcription factors of different classes produce mRNAs that do not encode for the DNA-binding domains across *Metazoa*_____159

Figure 5.1 *HoxA/D* genes share a large number of conserved 3'UTR motifs_____176

Figure 5.2 Hierarchichal clustering of *HoxA/D* genes based on shared 3'UTR motifs_____179

Figure 5.3 Hierarchichal clustering of *HoxA/D* genes based on co-expression patterns in the developing forelimbbud_____182

Figure 5.4 Shared *HoxA/D* 3'UTR motifs significantly recapitulate dynamic *HoxA/D* co-expression patterns in the forelimb_____185

Figure 5.5 *HoxA/D* 3'UTRs contain numerous RBP-target motifs_____189

Figure 5.6 The 3'UTRs of phylogenetically unrelated genes share *cis*-motifs that significantly recapitulate their expression patterns in the *Mus musculus* hindbrain__195

Figure 5.7 The 3'UTRs of phylogenetically unrelated genes share *cis*-motifs that significantly recapitulate their expression patterns in the *C. elegans* germline_____198

List of Tables

Table 2.1 *PCR Primers……… ………………………………………………………………….83*

Table 3.1 *– Experimental techniques used in the validation of mammalian Hox-miRNA interactions (data retrieved from miRTarBase)…………………………..………………..…108*

Table 3.2*- 3’UTR targeting predictions of experimentally validated miRNA-Hox interactions using PITA……..………………………………………...……………………...112*

Table 5.1 *- A list of 32 genes with rhombomere-restricted expression in the Mus musculus hindbrain………………………………………………………………………..…193*

Abbreviations

A3SS	Alternative 3’ splice site
A5SS	Alternative 5’ splice site
A-P axis	Anterior-posterior axis
abd-A	abdominal-A
Abd-B	Abdominal-B
AFE	Alternative First Exon
ALE	Alternative Last Exon
ANT-C	Antennapedia-complex
Antp	Antennapedia
APA	Alternative cleavage and polyadenylation
AS	Alternative splicing
bx	Bithorax
BX-C	Bithorax-complex
CNS	Central nervous system
Dfd	Deformed
Dll	Distaless
DNA	Deoxyribonucleic acid
DBD	DNA-binding domain
d.p.c.	days *post coitum*
DRP	Differential RNA processing
elav	embryonic lethal abnormal vision
ftz	fushi-tarazu
HX	Hexapeptide

HD	Homeodomain
lab	labial
mRNA	messenger RNA
mIRNA	microRNA
PAS	Polyadenylation signal
pb	proboscipedia
PcG	Polycomb group
PG	Paralogue group
PRE	Polycomb responsive elements
PWM	Position Weight Matrix
RBP	RNA-binding protein
RNA	Ribonucleic acid
PCR	Polymerase chain reaction
RI	Retention of Introns
RT-PCR	Reverse transcription polymerase chain reaction
Scr	Sex combs reduced
SE	Skipped Exons
SPR	Subtree Pruning and Regrafting
t3UTR	Tandem 3'UTRs
tTSS	Tandem transcription start sites
Ubx	Ultrabithorax
UTR	Untranslated region

My own interest in the development of the fins of fishes was early raised to a pitch; but when I told a lady that I was writing my thesis on this subject, her reply was, "What earthly good are fins? I never eat them." To the layman such aberrations of taste are beyond comprehension. In fact, there is no easier way of holding up learning to public scorn and ridicule than to repeat the titles of Ph.D. theses.

R.G. Harrison

(Harrison 1937)

Chapter I

General Introduction

1.1 – Preface.

The question of how animal development is controlled at molecular level is a central aspect of modern biological research. *Hox* genes have been shown to underlie animal morphogenesis along the main axis of a number of organisms. More specifically, the expression and morphogenetic action of *Hox* genes is segmentally restricted along the anterior-posterior axis (A-P axis) of both arthropods, like *Drosophila melanogaster* and mammals like *Mus musculus* (Pearson et al. 2005). In the latter group, differential *Hox* gene expression also mediates the morphogenesis of the secondary axis of the limbs, a process that shares a number of characteristics with the patterning of the A-P axis, leading to the proposal that the morphogenetic action of *Hox* genes has been co-opted to pattern this novel character of the Vertebrate clade (Lonfat et al. 2014).

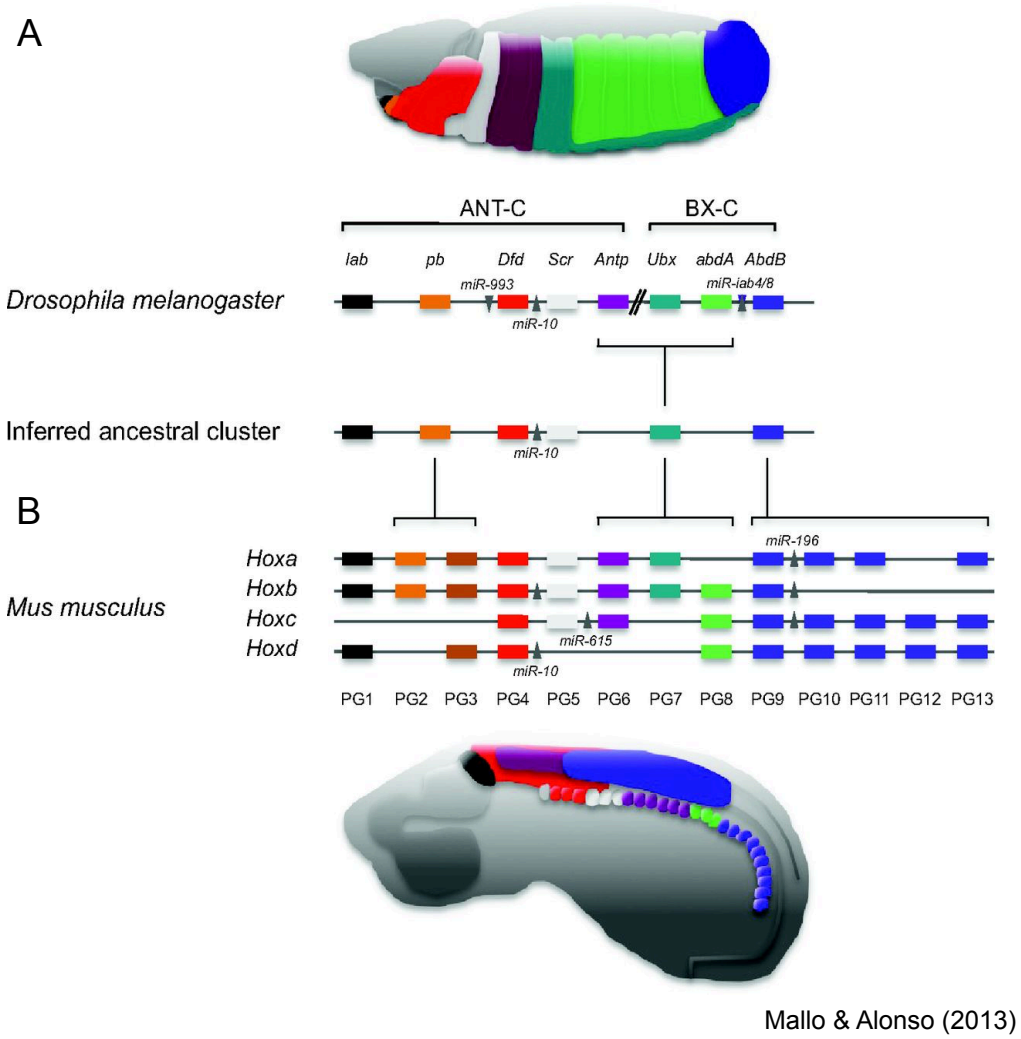
The precise regulation of *Hox* gene expression has been shown to impact the aforementioned morphogenetic effects of *Hox* products in both *Drosophila melanogaster* and *Mus musculus*. In particular, the control of *Hox* gene expression by chromatin and transcriptional regulation has been show to impact *Hox*-mediated developmental programs in both organisms, and in both the primary and secondary axis of mammals (Zakany & Duboule 2007; Mallo & Alonso 2013). The RNA-based regulation of *Hox* genes has a strong impact on *Hox* molecular patterns during the development of *Drosophila melanogaster*, via the production of alternative mRNAs from the same *Hox* locus, and their subsequent regulation by trans-acting factors. This level of regulation controls the expression patterns of at least half of the *Drosophila melanogaster Hox* genes (Thomsen et al. 2010; de Navas et al. 2011; Reed et al. 2010).

In the following sections of this Chapter, I explore how *Hox* genes, a group of related genes that display homology in sequence, molecular action and function across

metazoans, control the development of *Drosophila melanogaster* and *Mus musculus*, as well as the gene regulatory levels that help establish *Hox* expression patterns during the development of these organisms. I then look at the evolution of the *Hox* gene clusters in vertebrates, showing how the history of whole genome duplications in the vertebrate clade relates to *Hox* expression and function in mammals. With a focus on RNA-based regulation of *Hox* expression, I will then discuss development in an evolutionary context and introduce the specific aims of this thesis.

1.2 - *Hox* genes and the Homeodomain.

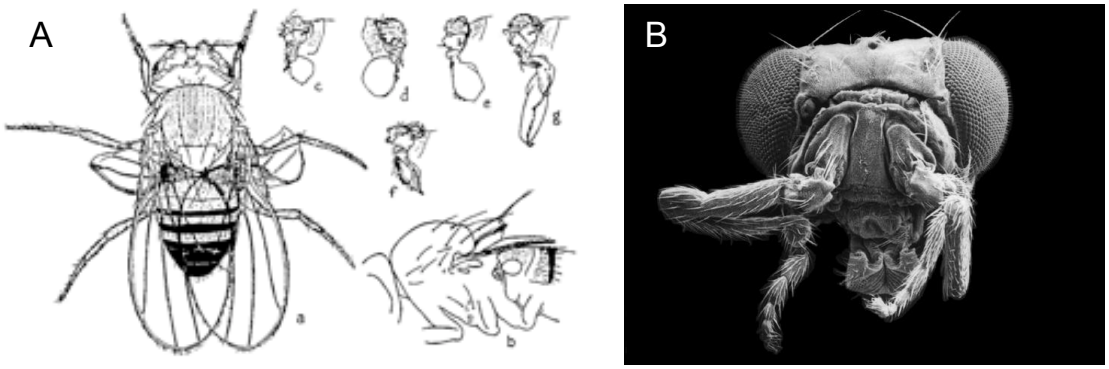
Hox genes usually occur in clusters within the genomes of animals, and encode for a family of evolutionarily related transcription factors that contain a DNA-binding domain, the Homeodomain (Pearson et al. 2005), (**Figure 1.1**). Hox transcription factors are present in most animals, being usually expressed along the anteroposterior axis of early embryos in a segmentally restricted manner (Pearson et al. 2005). Generally, the relative position of a *Hox* gene within a cluster is mirrored by the relative position of its segmental expression, a rule called *spatial colinearity* (**Figure 1.1A-B**). Through the differential effects on the transcription of target genes, the restricted expression of *Hox* genes in tandem segments confers differential identities to serially homonomous structures, leading to the differentiation of function and morphology that characterizes the anteroposterior axis of animals. This axis, as well as the morphological structures that characterize it in different animals (e.g. thoracic ribs and arms in mammals *versus* thoracic legs and wings in *Diptera*), characterizes the distinct body-



Mallo & Alonso (2013)

Figure 1.1 - *Hox* clusters are conserved across animals and confer positional information across the A-P axis during development (legend in the following page)**.**

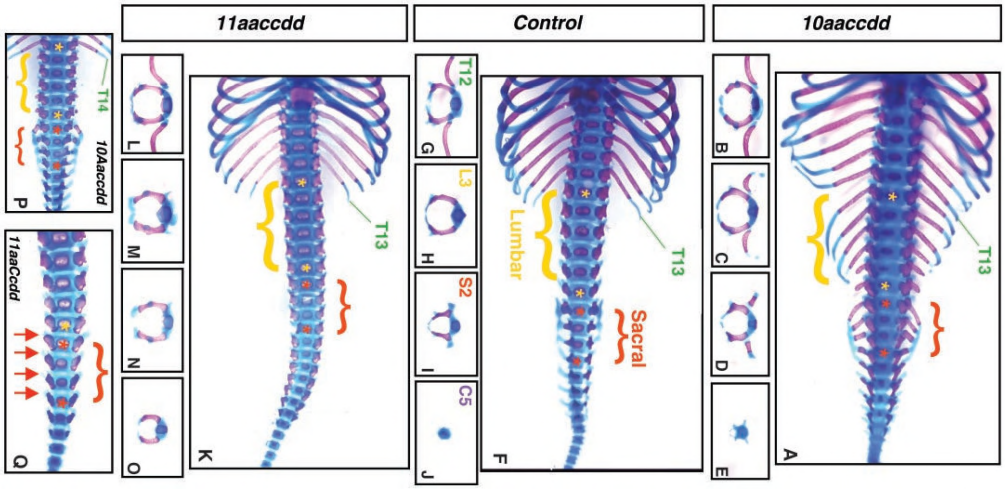
Figure 1.1 - *Hox* clusters are conserved across animals and confer positional information across the A-P axis during development. (A-B) Diagram depicting the genomic organization of *Hox* genes in *Mus musculus* (A) and *Drosophila melanogaster* (B), taken from (Mallo & Alonso 2013). (A) There are eight *Hox* genes clustered in two complexes in *Drosophila melanogaster*, the ANT-C and the BX-C complexes. The manner in which *Hox* genes are organized in the genome mirrors the embryonic gene expression patterns of *Hox* genes along the anterior-posterior (A-P) axis, a characteristic named spatial colinearity. More “anterior" genes like *Deformed* (*Dfd*) - part of the ANT-C - will be expressed in more anterior segments than *Hox* genes sitting in the other extremity of the cluster *e.g. Abd-B*. The *Drosophila Hox* clusters contain three miRNA *loci*, which encode for small RNA molecules that in some cases, like *miR-iab-4-5p/3p*, can target *Hox* mRNAs in a post-transcriptional manner. The *Hox* genes of *Drosophila* are hypothesized to descend from six *Hox* genes in the genome of the common ancestor of animals with bilateral symmetry (urbilaterian) (B) The *Mus musculus* genome has 39 *Hox* genes, organized in four clusters which sit in different chromosomes. The 39 *Hox* genes of mammals descend from a single *Hox* cluster by two rounds of genome duplication early in the vertebrate evolutionary lineage. As such, there are thirteen paralogue groups across mammalian *Hox* clusters. Within each paralogue group (PG), *Hox* genes have similar relative positions in different clusters and share sequence motifs, expression patterns and function. As with *Drosophila melanogaster* [(see (A)] the mammalian *Hox* clusters contain miRNA *loci* and display spatial colinearity between *Hox* genomic positions and embryonic axial expression patterns.



adapted from Bridges & Morgan (1923)

F. R. Turner, Indiana University

C



adapted from Wellik & Capecchi (2003)

Figure 1.2 - The mutation of *Hox* genes causes homeotic transformations in *Drosophila melanogaster* and *Mus musculus* (legend in the following page)**.**

Figure 1.2 - The mutation of *Hox* genes causes homeotic transformations in *Drosophila melanogaster* and *Mus musculus*. (A) Diagram depicting a homeotic phenotype due to a *bx* mutation in *Drosophila melanogaster* (Bridges & Morgan 1923). (A) In *bx* mutants, the third thoracic segment, which usually contains a pair of halteres, is transformed into the likeness of the T2 segment, showing a partial haltere-to-wing transformation. (**B)** Picture detailing the phenotype of an *Antp* gain of function mutation (F.R. Turner, Indiana University). This mutation leads to the ectopic expression of the *Hox* gene *Antennapedia* in the developing antennae of *Drosophila melanogaster*, leading to a antenna-to-leg transformation in this segment. **(C)** The mutation of *Hox* genes leads to homeotic transformations in *Mus musculus* (Wellik & Capecchi 2003). The mutation of *Hox* genes of paralogue groups 10 and 11 leads to homeotic transformation in the segments of the A-P axis in which they are developmentally expressed. The mutation of all copies of *Hox10* genes leads to a lumbar-to-thoracic homeotic transformation, in which ribs are formed in the lumbar region. Conversely, the mutation of *Hox11* copies in *Mus musculus* leads the extension of lumbar phenotypic fates toward the sacral regions of the axial skeleton.

plans (*baupläne*) of metazoans, and its diversification underlies major evolutionary transitions in the tree of life. As the expression of *Hox* genes is known to mediate the identity of these axial structures, *Hox* genes have been implicated in the evolution of the animal *bauplan*. The first line of evidence for these ideas involves the study of *Hox* mutations in which the morphology of one segment is transformed into the semblance of another. These transformations are usually titled *homeotic* (**Figure 1.2**).

William Bateson had coined the term *homeosis* in 1894 to broadly describe the types of morphological variation in which "something has been changed into the likeness of something else" (Bateson 1894). This effect had been recognized by Goethe and termed *metamorphy* 104 years previously, in his study of plant variation (Lewis 2004). As this is an ambiguous term, being also used to describe the process of metamorphosis, Bateson proposed a change in nomenclature so as to distinguish individual metamorphic variants within a population (*homeosis*) from the remaining metamorphic variation during ontogeny (*metamorphy* or *metamorphosis*) (Lewis 2004).

In 1915, Calvin B. Bridges and Thomas H Morgan discovered the first homeotic mutation in *Drosophila melanogaster* (Bridges & Morgan 1923) (**Figure 1.2A**). In their 1923 book (Bridges & Morgan 1923) these authors describe the discovery, by Calvin Bridges, of "a mutant eye-color like maroon" fly stock of which a male was crossed with a wild-type female. This led to a stock that lost this character in the F_2 generation; "However, approximately a quarter of the flies showed (culture 2203, September 22, 1915) a new character of a surprising nature (Bridges & Morgan 1923). These flies appeared to have two thoraxes with wings and bristles complete. (…) Some of these *bithorax* flies were mated together, and all the progeny were found to be bithoracic, constituting a pure-breeding stock of the recessive mutant." (Bridges & Morgan 1923). Wild-type *Drosophila melanogaster* flies usually have three thoracic segments, deemed

T1, T2 and T3, with the last as the most posterior. The T1 segment has a pair of ventral legs, as do the T2 and T3 segments. Additionally, T2 displays one dorsolateral pair of wings, while T3 has dorsolateral halteres, balancing organs that act as gyroscopes to control flight. In *bithorax (bx)* flies, Bridges and Morgan found that the T3 segment displays a modified morphology that resembles that of the immediately anterior segment, with both dorsal and ventral appendages resembling those of T2 (Bridges & Morgan 1923), (**Figure 1.2A**).

Subsequent forward genetic work in *Drosophila melanogaster* uncovered a number of additional homeotic transformations (**Figure 1.2B**). Interestingly, a number of these were mapped to two closely located regions in the right arm of the third chromosome of flies; these regions were deemed Antennapedia complex (ANT-C, (Kaufman et al. 1980)) and Bithorax complex (BX-C, (Lewis 1978)), (**Figure 1.2A-B**). The work of E.B. Lewis showed that the linear arrangement of mutant regions, inferred from classical genetic maps, was correlated with the location of their specific morphological effects along the anteroposterior axis, a phenomenon now known as spatial colinearity (Lewis 1978; Lewis 2004). This led to the proposal of a coordinate system of morphological patterning, in which the differential expression of *Hox* genes along the anteroposterior axis would lead to the differential identity and resulting morphology of segments along the same axis. In 1983, the innovative study of Bender and colleagues described the DNA sequence of 195 kilobases (kb) within the BX-C, providing a solid molecular basis for the study of Hox genes (Bender et al. 1983). This was followed by expression analyses that revealed that *Hox* genes were indeed expressed in discrete contiguous domains along the anteroposterior axis of the embryo, and that their order is collinear to the position of each respective *Hox* gene within the *Drosophila Hox* cluster (Mallo & Alonso 2013; Akam 1987; Harding et al. 1985),

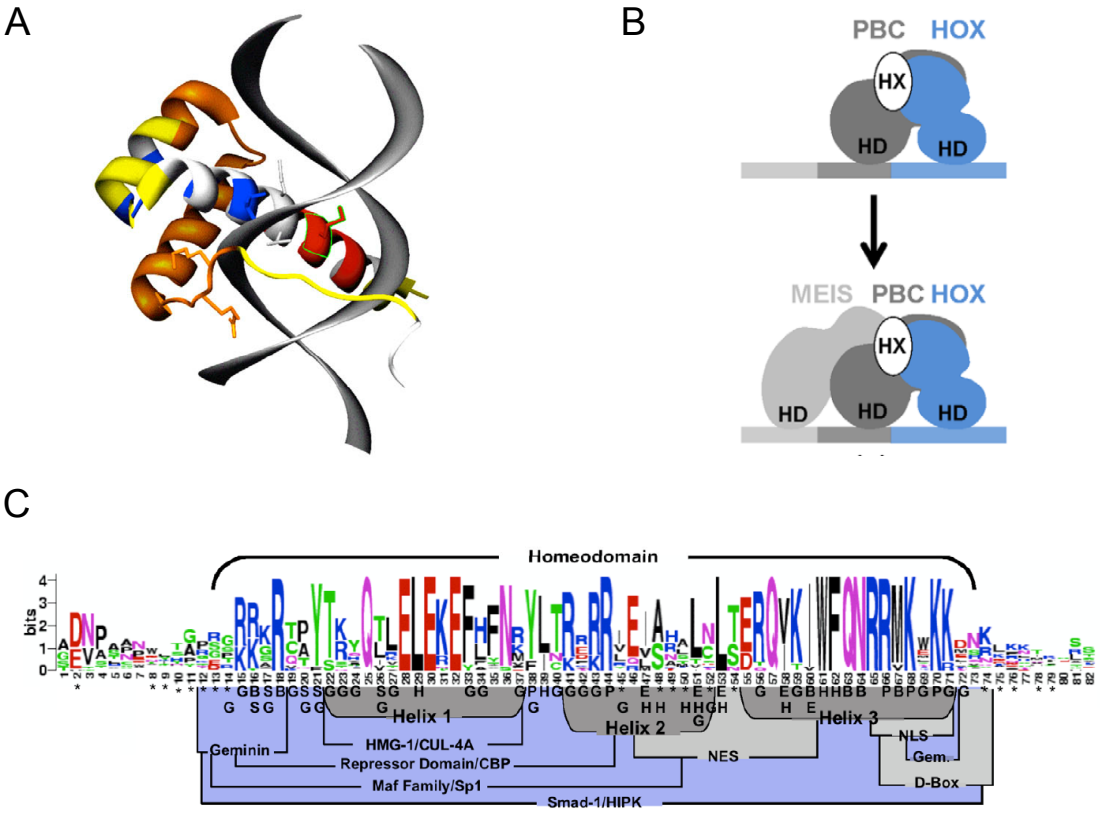
(**Figure 1.1A**).

Eight *Hox* genes have been described in *Drosophila melanogaster*. The 5 *Hox* genes with the most anterior embryonic expression - *labial* (*lab*), *proboscipedia* (*pb*), *Deformed* (*Dfd*), *Sex-combs reduced* (*Scr*) and *Antennapedia* (*Antp*) - are serially distributed within the ANTP-C, while the remaining 3 genes *Ultrabithorax* (*Ubx), abdominal-B (Abd-B)* and *abdominal-A (abd-A)* are located in the BX-C. The analysis of expression of *Hox* genes further confirmed that these were indeed serially expressed along the segments of the anteroposterior axis of fruitflies (Akam & Martinez-Arias 1985; Beachy et al. 1985; Harding et al. 1985; Karch et al. 1990; White & Wilcox 1984) (**Figure 1.1A**).

In 1984, Bill McGinnis (McGinnis et al. 1984) showed that upon Southern hybridization between an *Antp* cDNA probe (Carrasco et al. 1984) and genomic DNA of *Drosophila melanogaster,* not one but four non-contiguous complementary regions existed in a 100 kb stretch within the ANTP-C. This result indicated that different regions across the ANTP-C *Hox* complex shared specific sequences that were mapped to the transcription units of *Antp*, *Ubx* and *fushi tarazu* (*ftz*), a pair-rule gene found within the ANTP-C. Additionally, the authors of this study varied the hybridization stringency of the probes to show that more than 50 regions displayed some degree of hybridization to the probe used, indicating that this region is repeated outside of the *Hox* gene complexes. Independently, work by Scott and Weiner had shown that this shared sequence encoded for a protein domain with a high degree of homology (Scott & Weiner 1984). This observation led to the proposition that *Hox* genes share a close evolutionary history, and that their similarity in function can be attrituted to this protein domain – the Homeodomain (Scott & Weiner 1984; Desplan et al. 1988; Desplan et al. 1985).

All *Hox* genes contain a Homeobox sequence of around 180 base pairs in size, which encodes for a Homeodomain (**Figure 1.3**). The Homeodomain consists of a sequence of around 60 amino acids, and is found in the *Hox* genes of mammals, as well as in other Homeodomain-containing genes of Animals, Plants and Fungi (**Figure 1.3C**). When translated, this 60 amino acid sequence folds into three helices (1-3). Helices 2 and 3 form a helix-loop-helix structure that has the ability to bind DNA, activating or repressing transcription (Qian et al. 1989; Gehring 1993), (**Figure 1.3A, C**). This structure is similar to the helix-loop-helix domain of prokaryotic transcription factors (Gehring 1993). One of the two α-helices within the helix-loop-helix motif of the Homeodomain, helix 3 or the recognition helix, recognizes specific DNA sequences by insertion into the major groove of the DNA double helix (Qian et al. 1989; Gehring 1993; Affolter et al. 1990). In the case of the Antennapedia Homeodomain, the amino terminal region preceding the Homeodomain is flexibly disordered in solution and reaches into the minor DNA groove. An additional loop between helices 1 and 2 contacts the DNA backbone in the major groove (Qian et al. 1989; Gehring 1993; Affolter et al. 1990). These Homeodomain-DNA interactions are consistent with observations in other Homeodomain transcription factors (Kissinger et al. 1990; Gehring 1993; Wolberger et al. 1991). All these Homeodomain-DNA interactions are thought to stabilise the Homeodomain-DNA interaction (Gehring 1993). Indeed, the monomeric binding of the Antennapedia Homeodomain to DNA molecules was found to be highly specific as well as stable, with Antp-DNA complexes displaying an *in vitro* half-life of approximately 1.5 h (Affolter et al. 1990).

These observations led to the proposal that *Hox* genes act as regulators during the development of *Drosophila melanogaster*, binding operator DNA sequences in the vicinity of target genes, and subsequently influencing their transcription in a positive or



Adapted from Lynch *et al.* (2006) and Hudry *et al.* (2012)

Figure 1.3 – *Hox* genes contain the Homeobox, which encodes for the DNA-binding homeodomain, as well as other protein-protein interaction domains (legend in the following page)**.**

Figure 1.3 – *Hox* genes contain the Homeobox, which encodes for the DNA-binding homeodomain, as well as other protein-protein interaction domains. (A) Diagram depicting the structure of the Homeodomain (from (Lynch et al. 2006)). The Homeodomain is a helix-loop-helix DNA-binding domain. In its native protein conformation, the Homeodomain contains three α-helices. Helices 2 and 3 fold into a helix-loop-helix structure that has the ability to bind DNA, leading to the activation or repression of transcription. **(B)** Hox proteins bind DNA in a cooperative manner by interacting with PBC and MEIS-class factors, both Homeoproteins, in *Drosophila melanogaster* and *Mus musculus*. These interactions stabilise Hox-DNA molecular interactions. Hox proteins mediate their interaction with PBC factors through the hexapeptide, a short (I)YPWM(K) amino acid sequence that lies upstream of the Homeodomain. MEIS-class Homeoproteins also serve the molecular function of stabilising Hox-PBC-DNA interaction (from (Hudry et al. 2012)). **(C)** The Homeodomain consists of a sequence of around 60 amino acids, which fold into a characteristic helix-loop-helix structure (see panel A). This sequence is encoded by the 180 base-pair Homeobox, which is found in the *Hox* genes of *Drosophila melanogaster* and *Mus musculus*, as well as in other Homeodomain-containing genes of Animals, Plants and Fungi (from (Lynch et al. 2006)).

negative manner. The complement of *Hox* targets has since been shown to be numerous, as well as to differ between *Hox* genes (Pearson et al. 2005); this was shown to rely on the slightly different specificities of different Hox Homeodomains. For example, domain-swapping experiments in the *Drosophila melanogaster* Hox factor Antennapedia have show that functional specificity of the Antennapedia Hox factor resides partially on its Homeodomain, as the presence of an *Antp*-specific four amino acid region in the N-terminal region of the Homeodomain (RGQT) is sufficient to cause a homeotic transformation from antennae to legs upon ectopic overexpression (Furukubo-Tokunaga et al. 1993). In contrast, *Distalless (Dll)* was shown to be targeted by different *Hox* genes, being repressed by *Ubx*, *abd-A*, and *Abd-B* in the abdominal epidermis (Vachon et al. 1992; Pearson et al. 2005). This indicates that *Hox* genes can indeed share some targets (reviewed in (Sánchez-Herrero 2013)), although their specific effect can be antagonistic *e.g. Ubx* and *abd-A* either activate (*Ubx*) or repress (*abd-A*) the target gene *dpp* in the visceral mesoderm of developing *Drosophila* embryos.

The segmentally restricted morphological effects of *Hox* genes, as revealed by homeotic transformations upon *Hox* mis-expression, are thought to arise from the segmentally restricted expression of *Hox* genes. Hox protein products then bind DNA directly as monomers, *via* the Homeodomain. As these DNA stretches are small, ranging from 4-5 nucleotides, and are as such bound by different Hox proteins, further target specificity can be achieved by Hox heterodimerization with members of the two Homeodomain-containing *extradenticle* or *exd* (PBC family in mammals) and *homothorax* or *hth* (MEIS super-family in mammals), (**Figure 1.3B**). The Hox-Exd protein-protein interactions are mediated in part by the (I)YPWM(K) Hexapeptide (HX), a small domain that lies upstream of the Homeodomain (**Figure 1.3B**). As the same Hox protein has differential target-site specificities in monomeric or heterodimeric

contexts (Pearson et al. 2005), Hox interactions with co-factors are expected to further modify target-site choice. In both conformations, direct binding of *Hox* genes to DNA occurs *via* the Homeodomain and leads to activation or repression of specific targets. Given the large, noticeable effects of *Hox* mutations, this mechanistic model provides the first molecular mechanism for the control of animal *bauplan* development (Pearson et al. 2005). In the following sections, I will introduce the *Hox* clusters of vertebrates as well as their evolutionary history as it relates to the mammalian evolutionary lineage, and will explore the consequences of regulated *Hox* expression in the context of mammalian *bauplan* formation.

1.3 - The evolution of mammalian *Hox* clusters.

The discovery of serially shared sequences within the *Drosophila Hox* clusters led to subsequent investigations of their presence in the genomes of other animals. Using *Drosophila* cDNA probes for the Homeobox in low-stringency screenings of affinity to vertebrate DNA sequences, the first Homeobox-containing genes were discovered in *Xenopus* and the placental mammal *Mus musculus* (Favier & Dollé 1997; McGinnis et al. 1984; Carrasco et al. 1984). Like the *Hox* genes of *Drosophila melanogaster*, vertebrate *Hox* genes were found to lie in clusters and to display spatial colinearity between genomic organization and gene expression along the anteroposterior axis (Graham et al. 1989; Pearson et al. 2005). In conjunction with the homeotic effects of Hox mutations in *Drosophila melanogaster*, their extreme conservation in sequence and expression across animals led to the proposal of *Hox* genes as central determinants of animal *Baupläne*, and by definition, of their evolution.

Vertebrates are part of the chordates, a phylum that shares its deuterstosme mode of embryonic development, in which the blastopore of early embryos becomes the adult anus, with Echinoderms and Hemichordates. The last common ancestor of the Deuterostomes and the Protostomes, the evolutionary lineage leading up to *Drosophila melanogaster*, is usually referred to the Urbilaterian as it represents the most recent common ancestor of all animals with bilateral symmetry ("*Ur-*" is a German prefix that means "original"). Molecular estimates place the Deuterostome-Protostome split, and thus the existence of the Urbilaterian, at roughly 670 million years ago (Ayala & Rzhetsky 1998). As both Deuterostome and Protostome lineages have very similar *Hox* genes, it is generally considered that the most recent common ancestor of the *Drosophila* and Vertebrate *Hox* clusters also existed at around the same time. It is thus important to understand what happened to the *Hox* gene clusters since the 670 million between the *Drosophila melanogaster* and the Vertebrate lineages.

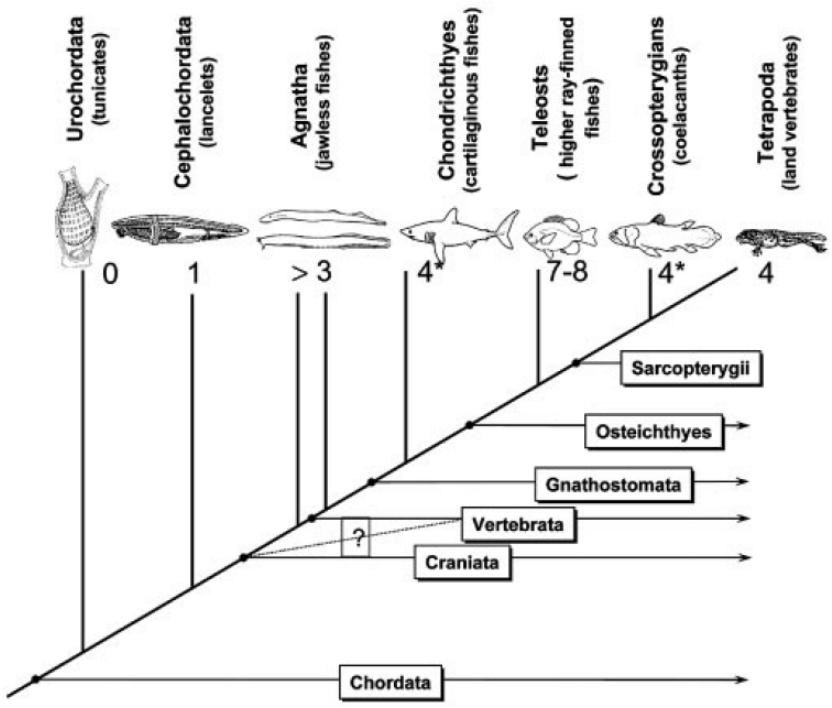
Unlike *Drosophila*, the *Hox* genes of vertebrates are found in four or more different genomic clusters. For instance, the *Hox* gene complement of amniotes (Reptiles, mammals and Birds) consists of 39 genes, unequally divided across four clusters that lie on four different chromosomes (Schughart et al. 1988; Ruddle et al. 1994) (**Figures 1.1B** and **1.4**)). Other *Sarcopterygii*, a vertebrate sub-clade that includes amniotes, as well as amphibians, coelacanths and lungfish, display a similar arrangement of *Hox* genes across four different clusters, with the exact number of *Hox* ranging from 38-42 genes (Liang et al. 2011), (**Figure 1.4B**). Other fish show an even bigger proliferation of *Hox* genes: the zebrafish *Danio rerio* has 47 *Hox* genes in 7 clusters (Meyer & Málaga-Trillo 1999) while the Atlantic salmon has a total of 118 *Hox* genes divided between 13 different genomic clusters (Mungpakdee et al. 2008).

Several authors have proposed that different rounds of genomic duplication underlie the various vertebrate expansions of the *Hox* repertoire (Ohno 1970) (Mungpakdee et al. 2008; Holland et al. 1994), (**Figure 1.4A**). Specifically, two rounds of whole-genome duplication are postulated at the base of the vertebrate lineage, after the Vertebrate-Urochordate split (2R), with a subsequent round of duplication (3R) occurring in the zebrafish evolutionary lineage (Ohno 1970), with the lineage of Salmonid fishes accumulating an additional duplication (4R) (Mungpakdee et al. 2008). After the two 2 rounds of whole-genome duplications that have occurred in the early Vertebrate lineage, most protein-coding genes were presumably lost as the protein-coding gene complements of most mammals is not the quadruple, but rather less than twice the size of that of the non-Vertebrate chordates (≈25.000 in humans *versus* ≈14.000 in *Amphioxus*, (Abbasi 2008)).

In stark contrast to these figures, the 39 *Hox* genes in mammals are orthologous to 13 *Hox* genes in *Amphioxus*, indicating that around 75% of *Hox* duplicates were retained in the mammalian lineage after the 2R duplication events. The Duplication-Degeneracy-Complementation model (DDC) provides an appropriate context to extend this observation in the present study (see **Chapter 3**). The DDC postulates that the retention of paralogues after gene duplication can be explained by, first, the subfunctionalization of each new paralogue, due to an accumulation of complementary loss-of-function mutations in sequences that encoded for different sub-functions in the ancestral gene (Prince et al. 2002). This leads, second, to the retention of both paralogues after gene duplication, as the molecular function of the ancestral gene can only be recapitulated if both paralogues are present (Prince et al. 2002).

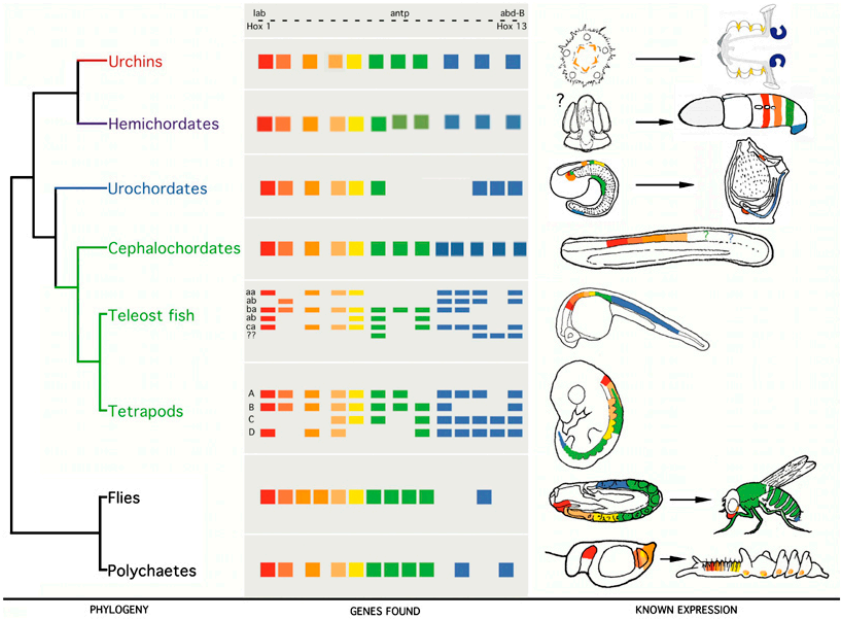
Individual *Hox* genes that are placed in the same relative position across different Vertebrate clusters were found to be very similar, sharing sequences as well as

A



Wagner *et al.* (2003)

B



Swalla (2006)

Figure 1.4 – The evolution of vertebrate Hox gene clusters by gene duplication (legend in the following page)**.**

Figure 1.4 – The evolution of vertebrate Hox gene clusters by gene duplication. (A) Numbers of *Hox* clusters in different clades of the chordate phylum (Wagner et al. 2003). Vertebrate genomes contain 4-8 *Hox* clusters. In cephalochordates and Urochordates, both sister-taxa to the Vertebrates, only 1 *Hox* cluster is found. This pattern results from that at least two rounds of whole-genome duplication at the base of the Vertebrate evolutionary lineage. In Teleosts, additional rounds of whole-genome duplications have resulted in the formation of more than 4 *Hox* clusters. The mammals *Mus musculus* and *Homo sapiens* are included in the Tetrapod clade, and have 4 genomic *Hox* clusters. **(B)** The organization of *Hox* genes within genomic clusters in the Bilateria (Swalla 2006). The two rounds of whole-genome duplication at the base of the Vertebrate lineage were followed by gene loss in the *Hox* clusters. In the Tetrapod lineage, this evolutionary history resulted in thirty-nine genes divided into four genomic clusters. *Hox* genes that occur in the same relative position within the *Hox* clusters of vertebrates are paralogous, sharing common ancestry by gene duplication. There are 13 groups of paralogous genes in vertebrate *Hox* clusters; these genes share specific sequences, relative genomic position, expression patterns and functions. The cephalochordates, a sister-clade to the Vertebrates, have 14 *Hox* genes organized in a single cluster. *Hox* genes 1-13 are orthologous of the paralogue groups 1-13 of Tetrapods. As such, the cephalochordates might present a *Hox* cluster that is similar to the one found in the chordate common ancestor.

patterns of expression, and occurring in the same relative position within each cluster (Pearson et al. 2005), (**Figures 1.1B** and **1.4B**). Given the evolutionary history of the vertebrate genome, and as mentioned previously, these commonalities across vertebrate *Hox* clusters are interpreted as representing *paralogy*, meaning that these similar *Hox* genes are paralogues of each other, a type of homology in which there is shared common ancestry due to gene duplication. As *Hox* paralogy occurs across the length of all four *Hox* clusters, the vertebrate *Hox* clusters are themselves paralogues of each other, and in conjunction, are orthologous (*i.e.* share homology due to speciation) to the single *Hox* cluster of most other animals. These observations provide an appropriate context to introduce the main genetic focus of our work, the *Hox* clusters of mammals.

As mentioned previously, mammals have 39 *Hox* genes that are divided across 4 genomic clusters, deemed A, B, C and D. In humans, these four clusters are located in different chromosomes: chromosomes 2, 7, 12 and 17 (Hokamp et al. 2003) (**Figure 1.1B**). Each mammalian *Hox* cluster has 9 to 11 *Hox* genes, which are paralogous to the *Hox1-13* genes of the cephalochordate *Amphioxus* (**Figure 1.4B**). An additional *Amphioxus Hox* gene, *Hox14*, was presumably lost after the Amphioxus/Vertebrate split, but before the genome duplications in the vertebrate lineage, as none of the four *Hox* clusters of mammals show evidence for the presence of a *Hox14* gene. In each mammalian cluster, *Hox* genes are conventionally termed by numbers 1-13, in ascending order of posteriority in expression. As such, each mammalian *Hox* gene has a coordinate defined by the cluster it is in, as well as its position within the cluster (e.g. the human *HoxA1* gene is part of the anteriorly-expressed *Hox* genes of cluster *A*) (Scott 1993). The aforementioned cross-cluster paralogy of mammalian *Hox* genes is thus characterized as such: two paralogous *Hox* genes are two *Hox* loci that are located in different *Hox* clusters, sharing similarity in sequence as well as within-cluster position.

As such, the number included in the name of each mammalian *Hox* gene also denotes its evolutionary relationship with its paralogues: the human *HoxA1* and *HoxB1 genes* are reciprocally paralogous, as are *HoxC13* and *HoxD13*.

Membership to a paralogue group can also be ascertained by analysing the combination of amino acid sequences that each *Hox* gene encodes. As mentioned before, all *Hox* genes share a Homeobox sequence that encodes for a Homeodomain. The Homeodomain is slightly different across PGs, but in only few cases can it offer enough resolution to ascribe paralogue group membership (Sharkey et al. 1997). As such most PG-specific sequences lie outside of the Homeodomain. Mammalian *Hox* genes from PGs 1-8 contain the hexapeptide motif (also present in *Drosophila Hox* genes, see previous section). The hexapeptide mediates interactions between *Hox* products of paralogue groups 1-8 and PBC-class homeoproteins. Members of paralogue groups 9 and 10 have a degenerate HX, consisting of a single tryptophan residue that nevertheless has been shown to also mediate the formation of Hox-Pbx1 DNA-binding complexes (Chang et al. 1996; Shen et al. 1996), (see **Figure 1.3B**). Members of PGs 11-13 do not exhibit protein-protein interactions with PBC factors, interacting instead with MEIS Homeoproteins. These molecular partners, however, are not mutually exclusive, as *Hox9 and Hox10* products can interact with members of both Homeoprotein classes.

Due to the individual pre-duplication evolutionary history of each vertebrate *Hox1-13* gene, paralogous *Hox* genes also share sequences with each other that are absent in other paralogue groups (PGs). For instance, Hoxa9-MEIS interactions were shown to rely on the first 61 amino acids of the Hoxa9 protein sequence (MIM, or MEIS interacting motif), a *Hox9* paralogue-group specific stretch of sequence that lies outside of the Homeodomain. The mammalian *Hox10* genes (*Hoxa10, Hoxc10,*

Hoxd10) are part of the posterior class of *Hox* genes, and share specific sequences like the M1 and M2 motifs that surround the Homeobox (Guerreiro et al. 2012). Even though *Hox9* genes display sequences that are somewhat similar to M1 and M2 in these regions, these are diagnostic of *Hox10* genes (Guerreiro et al. 2012). These PG-specific motifs are thought to reflect PG-specific molecular functions (Sharkey et al. 1997). In the case of *Hox10* genes, both M1 and M2 motifs underlie the repression of rib fates in the lumbar region of developing *Mus musculus* (Guerreiro et al. 2012). This repressive effect was shown to rely on the phosphorylation of Serine and Threonine residues within the M1 sequence (Guerreiro et al. 2012). As such, in this case, paralogue-specific domains exist that reflect paralogue-specific functions. Each paralogue-specific protein motif complement is thus combinatorial, and reflects the deep history of each PG.

In the next section, I describe how the regulated expression of mammalian *Hox* genes influences the embryonic development of mammalian structures along the primary anteroposterior axis, as well as the secondary proximal-distal axis of limbs.

1.4 - The impact of *Hox* expression on mammalian development.

During the development of mammals, groups of paralogous genes are activated sequentially along an axis, with the members of the paralogue groups *Hox1* and *Hox2* having an earlier and more anterior onset of embryonic expression than *Hox3* and *Hox4* genes, which in turn have earlier expression onsets and more anterior expression pattern than the remaining *Hox5-13* genes. As such, genes of the *Hox10-13* PGs are deployed late in mammalian development, and pattern more posterior structures (Wellik & Capecchi 2003). Unlike spatial colinearity, this aspect of *Hox* expression, known as

temporal colinearity, is not observed in other phyla (Mallo & Alonso 2013).

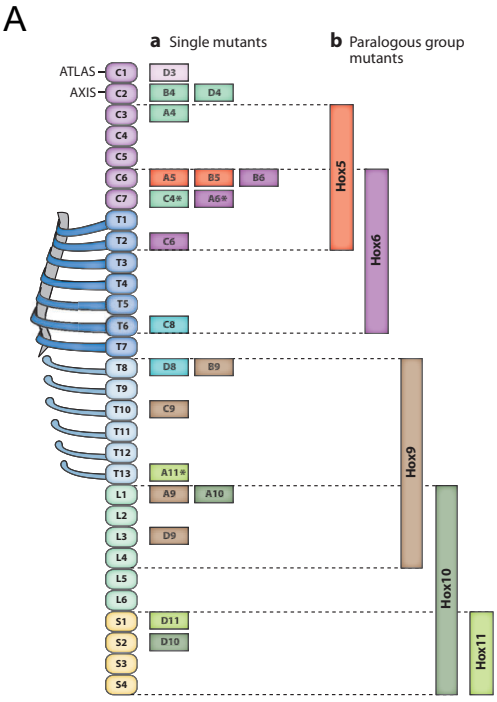
As with *Drosophila,* mutations in *Hox* genes introduce homeotic transformations along the anteroposterior axis of mammals, indicating that *Hox* genes also elicit segment-specific identities along this axis (Wellik & Capecchi 2003), (**Figure 1.5**). The mammalian *Hox* genes are involved in the patterning and differentiation of a number of axial structures. In this section, I will explore the developmental roles of mammalian *Hox* genes, focusing on the *Hox* involvement in patterning of the embryonic axial skeleton, the hindbrain and the limb of mammals (see **Figure 1.5**).

The axial skeleton of vertebrates develops, with the exception of the sternum (Wellik 2007), from transient mesodermal structures called *somites*, which are produced in a sequential manner beginning in the anterior end of the embryonic anteroposterior axis, at both sides of the neural tube. The sequential addition of somites occurs by epithelialization of mesenchymal cells at the anterior end of the presomitic mesoderm, and displays a precise periodicity that seems to be species-specific, occurring every 30 minutes in zebrafish, every 90 minutes in Chicken, and every 120 minutes in *Mus musculus*. In humans, this process is thought to occur at 20-35 days after conception, with each somite being formed in every 4–6 hours (Turnpenny et al. 2007). Molecularly, the periodicity in somitogenesis has been shown to rely, at least in part, on the oscillatory expression of a number of genes (Dale & Pourquie 2000), as well as in the establishment of a determination front, in which the anterior portion of a forming somite becomes progressively more exposed to a retinoic acid gradient (anterior to posterior) as it becomes less exposed to Wnt/Fgf signaling (established posteriorly). This in turn induces changes in gene expression that lead to segmentation at the anterior somitic border.

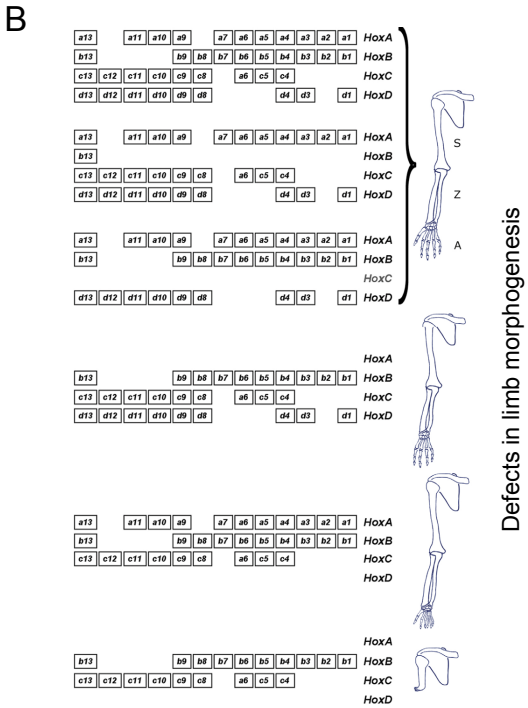
Although somites are serially homologous structures and thus morphologically

similar, they differentiate into distinct, segment-specific structures along the axial skeleton of mammals. The portion of the somite that forms the vertebrae and the rib cartilage of the axial skeleton - the sclerotome – migrates medially and dorsally towards the neural tube and fuse around it to form each vertebra. In the thoracic region, this process in prolonged by further lateral migration of slerotome cells to form the cartilage of ribs. In the lumbar region, however, this process and the consequent the formation of ribs, is absent (Wellik & Capecchi 2003).

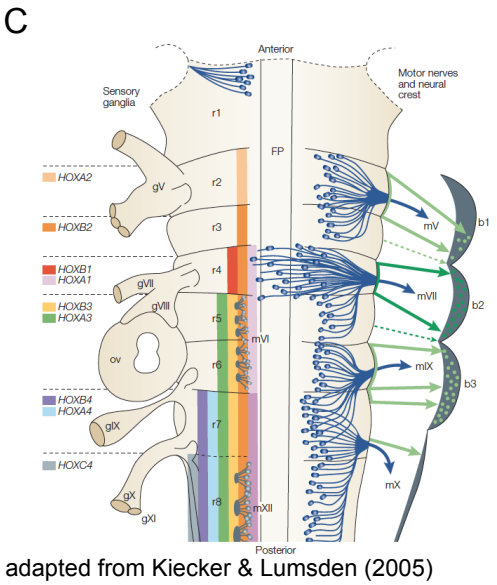
Hox genes exert morphogenetic control during the development of the axial skeleton at the presomitic mesoderm level. The anterior borders of expression of *Hox* genes are progressively established, following the spatial colinearity rule, and are fixed at 12 d.p.c. in the neural tube and presomitic mesoderm (Wellik 2007). The anterior



adapted from Alexander *et al.* (2009)



adapted from Zakany & Duboule (2007)



adapted from Kiecker & Lumsden (2005)

Figure 1.5 – Hox genes control the morphogenesis of the axal skeleton, limbs and hindbrain of mammals (legend in the following page)**.**

Figure 1.5 – Hox genes control the morphogenesis of the axal skeleton, limbs and hindbrain of mammals. (A) Diagram depicting the regions that are affected by *Hox* gene mutations in the mammalian axial skeleton. Anterior is up. (Alexander et al. 2009). Single *Hox* mutations affect the morphogenesis of specific regions along the A-P axis of the mammalian skeleton (a). The anatomical locus of these effects is correlated with the position of *Hox* genes within the mammalian *Hox* clusters, a characteristic called "spatial colinearity". For example, the mutation of *Hox6* genes affects more anterior regions of the skeleton, while the mutation of *Hox10* genes leads to morphogenetic defects in the more posterior lumbar and sacral regions. When groups of *Hox* paralogues are mutated in *Mus musculus* (b), broader phenotypic defects are found, affecting whole regions of the axial skeleton. For example, the mutation of all *Hox10* paralogues affects lumbar and sacral regions, leading to a homeotic transformation (see Figure **1.2C** and (Wellik & Capecchi 2003)). **(B)** Diagram depicting the regions that are affected by *Hox* cluster deletions in the mammalian forelimb (Zakany & Duboule 2007). The deletion of all but one of the *HoxB* genes, as well as a whole *HoxC* cluster deletion, leads to limb phenotypes that are similar in anatomy to the wild-type forelimb in *Mus musculus*. The deletion of either *HoxA* or *HoxD* clusters, however, leads to clear phenotypes due to lack of *HoxA/D* expression in the developing limb. In the event of a double *HoxA/D* cluster conditional knockout, the forelimbs of *Mus musculus* show a severe phenotype, with the disappearance of the autopod and zeugopod regions, as well as a severe truncation of stylopod. (**C)** Diagram depicting *Hox* expression patterns in the vertebrate hindbrain (Kiecker & Lumsden 2005). Genes of the *Hox* PGs 1-4 are expressed during the development of the vertebrate hindbrain. *Hox* genes also show spatial colinearity between their genomic organization and expression patterns in the vertebrate hindbrain, *e.g. Hox1* genes have more anterior expression borders than *Hox4* genes. This transient tissue is initially unsegmented but becomes compartmentalized into eight rhombomeres at around 9.5 d.p.c. during *Mus musculus* development. Each rhombomere shows a specific combination of *Hox* gene expression, which in turn influences segment-specific motoneuron projections and neural crest cell migration.

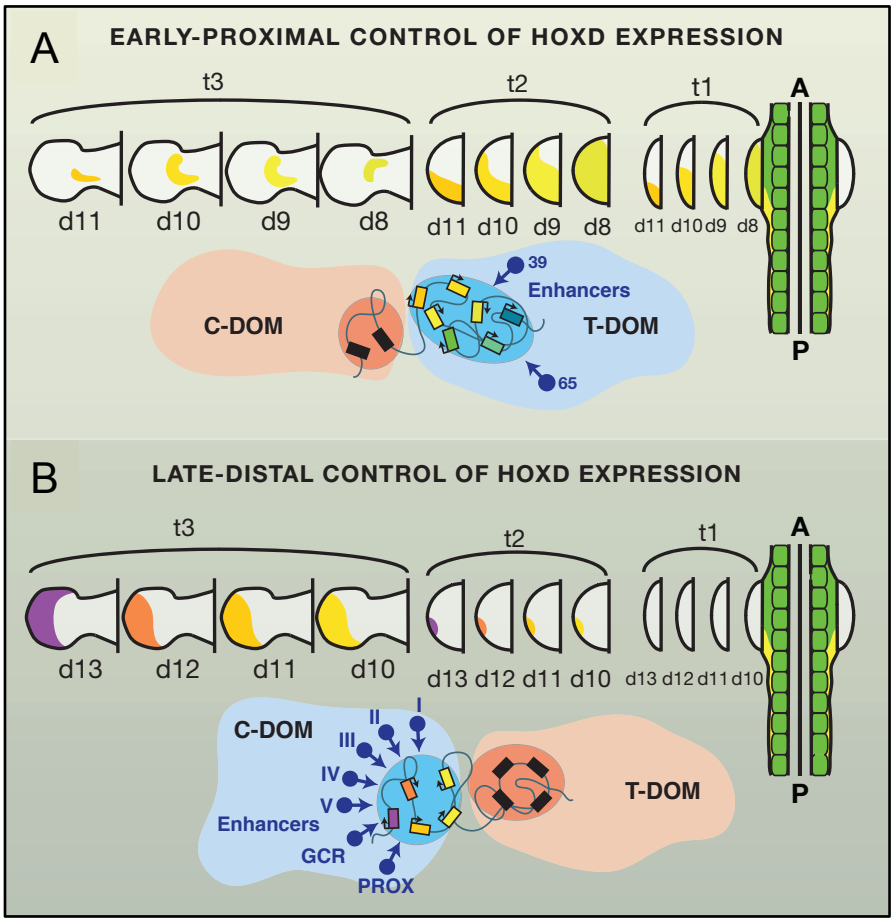
borders of *Hox* expression, as well as the specific combination of *Hox* input in each presumptive somite, are correlated will then confer an individual axial identity to the resulting skeletal structure (Wellik 2007), (**Figure 1.5A**). For example, the anterior border of *Hox10* expression in the presomitic mesoderm sits at the border between the presumptive thoracic and lumbar regions (Wellik 2007; Wellik & Capecchi 2003) (see **Figure 1.5A**). The mutation of all *Hox10* paralogues (*Hoxa10*, *Hoxc10* and *Hoxd10*) of *Mus musculus*, leads to an individual with no lumbar vertebrae; instead, ribs progress posteriorly, beyond the lumbar-sacral border (Wellik & Capecchi 2003), (see **Figure 1.2C**). In the case of a *Hox11* null (*Hoxa11*, *Hoxc11* and *Hoxd11*), lumbar vertebrae are established in sacral regions (Wellik 2007), (**Figures 1.2C** and **1.5A)**. Crucially, at least five of the six alleles of each paralogue group must be mutated, in both cases, for the phenotype to become apparent, highlighting the great degree of redundancy and co-expression of mammalian *Hox* paralogues (**Figure 1.2C**).

Similarly, *Hox* also pattern the developing brain of mammals along the anteroposterior axis (**Figure 1.5C**). The hindbrain, or rhombencephalon, is the terminal part of the developing vertebrate brain. Initially a featureless structure, it becomes segmented along the anteroposterior (AP) axis into eight compartments called rhombomeres (r) (see **Figure 1.5C**). These anatomical units are formed though the complementary expression of the ligand molecules Ephrins (expressed in r2, r4 and r6) and their Ephrin (Eph) tyrosine kinase receptors (expressed in r3 and r5, reviewed in (Dodelet & Pasquale 2000; Alexander et al. 2009)). Like their associated receptors, Ephrins are membrane-bound and can transduce extracellular signals through the intracellular phosphorylation of tyrosine residues following receptor binding (Dodelet & Pasquale 2000). In the hindbrain, this process gives rise to a bidirectional signalling cascade that alternates cell-cell adhesion properties along the AP-axis and consequently

sorts cells with different adhesion properties into adjacent rhombomeres. Each rhombomere then undergoes patterning and differentiation, influenced by the rhombomere-specific combinatorial expression of *Hox* genes (PGs 1-4), and gives rise to different morphological features that perdure in the adult brain, like the facial motor nerve root in rhombomere 4, and the vagus nerve root in rhombomere 7 (reviewed in (Kiecker & Lumsden 2005) and (Alexander et al. 2009)). For most *Hox* genes, the anterior border of expression coincides with rhombomere boundaries and extends posteriorly across rhombomeric compartments in the hindbrain (Alexander et al. 2009). This is not the case for Hoxb1, whose expression is restricted to rhombomere 4 (Alexander et al. 2009).

Hox expression also mediates the axial patterning of the developing limbs of mammals (**Figure 1.5B**). Here, I will focus on the main effects of *Hoxa*/*Hoxd* genes on limb morphogenesis, as whole-cluster deletions of *Hoxc* and *Hoxd* genes have little effect on adult limb phenotypes, while a *Hoxa*/*Hoxd* double deletion abolishes the morphogenesis of the limb (Zakany & Duboule 2007). I will also focus on the development of the forelimb, as more data are available for the manner in which *Hox* genes influence the morphogenesis of this organ (**Figure 1.5B**).

Briefly, the morphogenesis of *Mus musculus* forelimbs initiates with the formation of a sub-ectodermic bulge - the limb bud – at both sides of the trunk. This bud is formed by mesenchymal cells derived from the lateral plate and somitic mesoderms, and appears at a precise position that is determined by retinoic acid, Tbx5 and *Hox* gene expression in the trunk. The limb bud is then patterned as it proliferates posteriorly. As with axial skeleton and hindbrain development, the forelimbs are segmented structures: the proximal regions of the limb bud give rise to the adult humerus (stylopod), while the distal regions become the adult wrist and digits (autopod)



adapted from Andrey & Duboule (2014)

Figure 1.6 – HoxD genes are subjected to chromatin and transcriptional regulation in the developing mammalian limb (legend in the following page).

Figure 1.6 – *HoxD* genes are subjected to chromatin and transcriptional regulation in the developing mammalian limb. (A-B) Diagram detailing the control of early and late phases of *HoxD* expression during the development of the mammalian forelimb (from (Andrey & Duboule 2014)). (**A)** The expression of *Hox* genes of the *HoxD* cluster is initially restricted to posterior regions of the forelimb bud, becoming progressively distal as forelimb development progresses. At this stage *HoxD* expression is controlled at the chromatin and transcriptional levels. Proximal *HoxD* genes (*Hoxd8-d11*), which exhibit accessible chromatin states at this stage, directly contact Telomeric enhancers, leading to early *Hoxd8-11* transcriptional activation (T-DOM). **(B)** In later stages of forelimb bud *HoxD* expression, the posterior *HoxD* genes (*Hoxd12*-13) become transcriptionally active, and *Hoxd10-11* genes maintain active transcriptional states. *Hoxd8-9* genes, however, do not exhibit limb bud expression. The expression of *HoxD* genes is controlled by another regulatory environment (C-DOM) at this stage, involving both chromatin states and transcriptional activation. Here, centromeric enhancers control the activation of more posterior *Hox* genes. The previously active T-DOM regulatory landscape in silent at this stage, leading to the transcriptional inactivation of proximal *HoxD* genes. This mechanism is proposed to underlie the temporal colinearity of *Hox* gene expression during mammalian forelimb bud development.

(**Figure 1.5B**). The central region becomes the ulna and radius (zeugopod), which connect the two structures (**Figure 1.5B**).

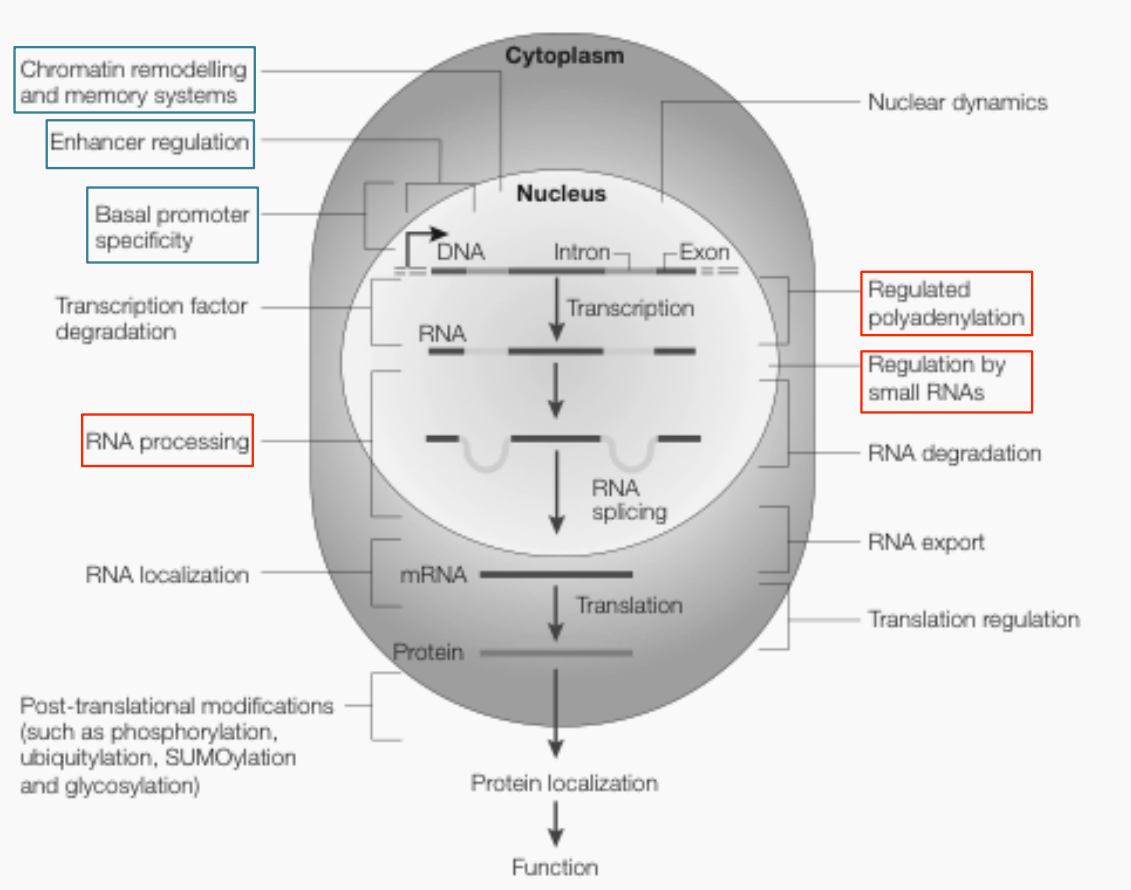
As with the trunk, *Hox* expression displays temporal and spatial colinearity during forelimb development (**Figure 1.6B**). The expression of both *HoxA* and *HoxD* genes is established in two waves. The first wave of expression establishes an asymmetrical distribution of *HoxA* and *HoxD* expression along the A-P axis of the early limb bud, with genes of both clusters being expressed in proximal limb territories. In the second phase of expression, *HoxD* genes are then posteriorly restricted as development progresses, a nested or recursive strategy that also resembles *Hox* expression in the trunk (Zakany & Duboule 2007) (**Figure 1.6B**). The similarities in expression colinearity and nestedness led to the proposition that the mode of *Hox* expression in the limb was co-opted from the trunk-patterning *Hox* network (Zakany & Duboule 2007). In *HoxA* genes, however, this later phase of limb patterning does not involve a progressively distal restriction of gene expression, leading authors to propose that in this second phase of gene expression, *HoxA* and *HoxD* genes might be under the control of distinct regulatory mechanisms (Zakany & Duboule 2007).

As previously mentioned for other contexts, *Hox* mutations lead to defects in the adult forelimb that correlate with the developmental expression domains of these genes (**Figure 1.5B**). For instance, the simultaneous mutation of *Hoxa13* and *Hoxd13* leads to the near absence of the adult autopod, while compound mutants for *Hoxa11* and *Hoxd11* have a severely shortened zeugopod.

1.5 – The regulation of mammalian Hox expression at the chromatin and transcriptional levels.

Up to this point, most of this chapter consists in the description of the genomic organization of *Hox* genes and its evolution, as well as the broad aspects of *Hox* expression in mammalian tissues, and its functional consequences. Here, I wish to start exploring the manner in which *Hox* gene expression is itself set up and, crucially, maintained, arguing that the study of the regulation of *Hox* expression, as well as its evolution, offers important insight for the evolution of gene regulation and its control of developmental programs. Our argument will follow the traditional view of hierarchical gene expression, providing illustrative examples of *Hox* regulation at the chromatin and transcriptional levels, and then focusing on RNA processing, post-transcriptional and post-translational levels of regulation. Concomitantly, I will provide examples as to how the regulation of *Hox* gene expression at specific regulatory levels mirrors that of other genes. In this section, I will focus on chromatin and transcriptional levels of *Hox* gene regulation, using genes of the *HoxD* cluster as illustrative examples.

Chromatin states are known to influence *Hox* gene expression in developing mammals (**Figures 1.6** and **1.7**). In early development, where *Hox* gene activity is usually silent, the chromatin in *Hox* clusters exhibits Histone marks that are consistent with an inactive state, and a consequent repression of gene expression. These consist of high levels of trimethylation in lysine 27 of histone H3 (H3K27m3), correlated with inactive chromatin states, and low levels of trimethylation in lysine 4 of histone H3 (H3K4m3) (Mallo & Alonso 2013). Broad changes from H3K27m3 to H3K4m3 marks that temporally correlate with the activation of *Hox* gene expression have been observed in the early development of the mouse tail, where the activation of *Hox* expression is followed by a strong increase in H3K4m3 and concomitant depletion in H3K27 trimethylation across the *HoxD* cluster (Soshnikova & Duboule 2009; Mallo & Alonso 2013). Similarly, the development of the limb has been show to exhibit a similar



adapted from Alonso & Wilkins (2005)

Figure 1.7 – *Hox* gene expression is subjected to a host of regulatory levels. Diagram depicting the several levels of *Hox* gene regulation (adapted from (Alonso & Wilkins 2005)). During the development of both *Drosophila melanogaster* and *Mus musculus*, *Hox* gene expression is regulated at the chromatin and transcriptional levels (blue). In *Drosophila melanogaster*, additional levels have been shown to further shape *Hox* expression in the developing embryo. Among these is the regulation of differential mRNA processing by alternative splicing and regulated polyadenylation, as well as post-transcriptional regulation by small RNAs (like miRNAs, red). Although the latter regulatory level has been shown to impact *Hox* expression during mammalian development (Hornstein et al. 2005), the extent to which miRNA-based regulation and differential RNA processing impact *Hox* gene expression patterns during the development of mammals remains largely unexplored.

correlation in *Hox* chromatin changes (**Figure 1.6**). In the aforementioned late phase of *HoxD* expression in the forelimb, where the initially posterior pattern of *Hox* expression becomes progressively distal, *HoxD* genes of distal posterior cells display a loss of H3K27me3 histone modifications, and show chromatin de-compaction, when compared to *HoxD* clusters in the anterior domain (Williamson et al. 2012) (**Figure 1.6B**). This chromatin-level regulation of *HoxD* genes has been shown to rely on the activity of Polycomb-group (PcG) proteins (Williamson et al. 2012).

PcG proteins are been known to mediate *Hox* repression during *Drosophila melanogaster* development (Lewis 1978; Mallo & Alonso 2013). In fruitflies, proteins of this complex bind to *cis*-regulatory regions deemed polycomb responsive elements (PREs) (Mallo & Alonso 2013), and exert their repressive functions across large stretches of DNA that include more than one *Hox* gene. Indeed, both ANT and BX *Hox* complexes seem to be included in the same polycomb-responsive unit, indicating that this is a broad level of *Hox* regulation (reviewed in (Mallo & Alonso 2013)). In vertebrates, however, the mechanistic involvement of PcGs on *Hox* expression remains largely unresolved; in the case of *Hoxd11 Hoxd12*, it seems to rely on the *homing* of PcG proteins to regions that resemble *Drosophila* PREs, as they are able to repress the expression of a reporter when in a genomic context (Woo et al. 2010).

In *Drosophila melanogaster*, *Hox* transcriptional input relies on a cellular memory system that is set-up and maintained by PcG proteins, as well as the control of transcription by earlier segmentation genes and other *Hox* products (Mallo & Alonso 2013). In the latter example, a phenomenon known as *posterior prevalence* occurs, in which posteriorly expressed *Hox* genes are able to repress the expression of more anterior *Hox*, effectively achieving contiguous *Hox* expression. This phenomenon helps establish the segment-specific expression of *Hox* genes in *Drosophila*, having thus key

functional consequences in the restriction of morphogenetic action of *Hox* genes to specific embryonic compartments (Mallo & Alonso 2013).

The control of *Hox* gene expression has also been shown to rely on regulated transcription in the context of mammalian development. The developing forelimb of *Mus musculus* offers us, once again, an illustrative example of this level of *Hox* regulation, as it has been shown that a switch in *HoxD* transcriptional regulation underlies *Hox* temporal colinearity in the developing limb. In the early phase of limb budding, *Hoxd8-Hoxd11* genes respond to a Telomeric transcriptional domain (T-DOM), being under the control of Telomeric enhancers that trigger *HoxD* activation (**Figure 1.6**). *Hoxd12* and *Hoxd13* genes do not seem to respond to this regulatory landscape. In a second phase of transcriptional activation, *Hoxd8* transcription is silent, as is T-DOM mediated *Hox* activation is lost. However, more *posterior* genes like *Hoxd12* and *Hoxd13* are transcriptionally activate due to the action of Centromeric enhancers, which are located in a 3D nuclear regulatory domain deemed C-DOM. At early stages, C-DOM is inactive; conversely, T-DOM is inactive at later stages of limb development. More central genes, like *Hoxd9*, *Hoxd10* and *Hoxd11*, appear to respond to both regulatory domains. In the context of limb development, it seems, therefore, that different sets of long-range enhancers, lying in distinct regulatory centres, control the differential temporal expression of proximal and distal *Hox* genes by means of a transcriptional regulation switch.

Recently, Lonfat and colleagues (Lonfat et al. 2014) have demonstrated that similar enhancers mediate the activation of *Hox* genes in the mammalian development of both distal limb regions and external genitalia. Although the authors show that there are transcriptional enhancers which are specific to one or another context, the shared enhancers support the idea that *Hox* transcriptional regulation in limb development was

co-opted from an older *Hox* cascade involved in the patterning of the primary developmental axis of animals. Additionally, *HoxA* genes show chromatin organization that is comparable to *HoxD*. Both clusters came into existence after the first round of genome duplication in the evolutionary lineage that leads to vertebrates (see above), which occurred in an invertebrate ancestor; *Hox* clusters C and B are though to have derived from clusters A and D after the second round of duplication. As such, the authors propose that similar chromatin conformations in the *HoxA*/*HoxD* clusters reflect old "regulatory topologies", which could have facilitated the recruitment of similar transcription factors to *Hox* genes. The authors advance that these ancient regulatory topologies "may have both favoured and constrained the evolution of pleiotropy" of *Hox* genes (Lonfat et al. 2014).

While the appropriate locus of *Hox* expression seems to rely on chromatin and transcriptional regulation, there is accumulating evidence that RNA-level regulation strongly impacts the establishment and quality of *Hox* expression in mammals. In the following section, I review evidence that supports a role for the RNA-based regulation of *Hox* genes in mammals.

1.6 – RNA-processing in mammalian *Hox* genes.

In this section, I argue that although chromatin and transcriptional regulation are key steps in the establishment of *Hox* gene expression during mammalian development, these regulatory levels might offer an insufficient explanation for the subsequent morphogenetic action of *Hox* genes. More, I sustain that once precise chromatin and transcriptional inputs result in the activation of *Hox* gene expression, a mature *Hox* RNA (mRNA) is subjected to a number of subsequent levels that impact the final output

of *Hox loci*. To this end I offer, first, examples of RNA-based *Hox* regulation in invertebrates, and submit, second, that there is mounting evidence for the involvement of RNA-level regulation in the establishment of precise mammalian *Hox* gene expression as well.

In eukaryotes, RNA processing is a gene regulatory level that generates a mature RNA copy from a certain DNA template. This regulatory level involves the coordination of different mechanisms of gene expression regulation. In this section, I will restrict our argument to genes under the transcriptional control of RNA Polymerase II, which includes all eukaryotic protein-coding genes, as well as most microRNAs (miRNAS). First, the RNA Polymerase II holoenzyme will start producing an RNA copy of a gene at a specific transcription start-site. The precise choice of transcription site is influenced by the composition of a proximal core promoter, as well as enhancer sequences to which *trans*-acting factors are bound. This level of regulation defines the 5'-end start of an RNA molecule, leads to the 5'-capping of the first ribonucleotide of an mRNA and is thus the first step in RNA processing. Crucially, this process can be regulated to elicit alternative transcription start-sites (TSSs). It is well known that many vertebrate and invertebrate genes (including *Hox* genes like the *Drosophila Antp locus*) display alternative promoter usage, leading to alternative first exons (AFEs). However, alternative TSSs may be also under the control of the same promoter, adding another layer to regulation of transcriptional initiation. For example, TSS usage changes dynamically with the transcriptional activation of the zygotic genome during the maternal-to-zygotic transition in zebrafish (*Danio rerio*) (Haberle et al. 2014). During this process, the usage of a maternal TSS that requires an A/T-rich motif is replaced by a zygotic TSS *grammar*, which exhibits a dependency on less specific motifs (Haberle et al. 2014). The alternative TSS *grammars* often coexist physically in the same

vertebrate core promoters, indicating that TSS choice involves a mechanism other than mere alternative promoter choice (Haberle et al. 2014). Tandem transcription start-sites (tTSSs) are widespread in mammalian genomes (Kawaji et al. 2006). Although the mechanisms are less clear in this case, some authors have shown that the usage of alternative TSSs within core mammalian promoters is associated with CpG islands, epigenetic imprinting, and multimodal promoters, which may have distinct modular sequences that underlie the alternative TSS choice, as with *Danio rerio* (Kawaji et al. 2006; Nepal et al. 2013). Crucially, these authors found evidence for a tissue specific TSS selection, indicating that this RNA processing level might be key in the establishment of tissue-specific patterns of expression. While alternative transcription start-site choice might influence the composition of an mRNA's open reading-frame - the stretch of RNA sequence that can be translated into a protein, it necessarily affects the length and composition of 5'untranslated regions (5'UTRs). In a recent study, Shifeng Xue and colleagues have found that *Hoxa9* 5'UTR sequences exhibit two interesting sequences, an internal ribosomal entry-site (IRES) that bypasses the usual 5'-cap dependent translation, as well as a translation inhibitory element (TIE), which inhibits cap-dependent mRNA translation (Xue et al. 2015). The authors then show that these sequences are a *sine qua non* condition for the translation of *Hoxa9* in the developing axial skeleton, and that the removal of these sequences leads to a homeotic transformation in which the rib-forming T13 vertebra assumes a lumbar fate (Xue et al. 2015). The authors also show that other *Hox* mRNAs, like *Hoxa4*, *Hoxa5* and *Hoxa11* also display IRESs sequences, and that these control the translation of a reporter construct (Merritt et al. 2008; Kondrashov et al. 2011). These IRESs may be conserved in *Drosophila*, while the TIE module is absent in zebrafish and amphibians, indicating that it might consist of an evolutionary novelty in the mammalian lineage, in which *Hox*

outputs are greatly impacted at an RNA-level regulation (Xue et al. 2015). These data suggest that the regulation of alternative transcription start site choice in mammalian promoters can influence the composition of a *Hox* mRNA, affecting its subsequent regulation and eventually function.

Once transcriptional initiation is successful and a nascent pre-mRNA is capped, the RNA Polymerase II holoenzyme proceeds transcription in the 5'-3' direction along the template gene. As it does this, it creates a 3'-5' pre-mRNA template of the whole gene's DNA sequence. eukaryotic genes, however, are discontinuous, in that they present stretches of intragenic sequences – introns – that are not present in mature RNAs. As such, for the formation of mRNAs that encode for coherent protein sequences, introns need to be excised from pre-mRNAs, with a concomitant joining of protein-coding stretches, or *exons*; this process is called *splicing*, and it typically generates coherent RNA ORFs. Some introns are constitutive, meaning that they are always excised from a pre-mRNA. All 39 mammalian *Hox* genes have at least one such intron. This kind of RNA processing operation relies on one of the most complex known molecular ensembles, the spliceosome (Nilsen 2003).

The spliceosome is a ribonucleoprotein complex that is similar, in size, to the ribosome (Nilsen 2003), and may include up to 300 distinct proteins in some contexts (Nilsen 2003). This complex is co-transcriptionally active in the nucleus and includes small nuclear ribonucleoproteins (snRNPs) U1, U2 U4 U5 and U6. The U1 snRNP recognizes an intronic GU dinucleotide (the 5' or donor splice site), while the U2AF1 protein recognizes a downstream AG intronic sequence (the 3' or acceptor splice site). The splice sites inclusively define the boundaries of the intronic sequence that will be spliced out. Additionally, introns usually contain an adenine nucleotide (branch-point), to which the SF1 protein binds, and a polypyrimidine tract between the latter and the 3'

splice site, which is bound by U2AF2. The U2 snRNP then supersedes SF1, binding to the branch-point. This is followed by the binding of a U5, U4 and U6 trimer, with the first snRNP recognizing the exon at the 5' side and the U6 snRNP binding to U2. The U6 snRNP then displaces U1 at the 5' splice site. Upon release of U4, The U6/U2 complex, which had brought the 5' splice site and the branch-point together, catalyse a trans-esterification reaction in which the 5'end of the intron binds the adenine branch-point nucleotide, forming a lariat RNA structure. U5 then binds the 3' splice site and the 5' splice site is cleaved, releasing one side of the lariat from the pre-mRNA. While U2, U5 and U6 snRNPs remain bound to the half-released lariat, the 3' splice site is cleaved and the 5' and 3' exons are ligated in an ATP-dependent manner. Once all introns are spliced out, a pre-mRNA becomes an ORF-carrying mature RNA (mRNA) (Will & Lührmann 2011). This is the canonical splicing reaction, occurring in the overwhelming majority of eukaryotic protein-coding genes (Will & Lührmann 2011).

As with the choice of TSS, the exclusion of introns can be optional, and intronic excision may thus be subjected to regulation. In these cases, multivalent loci produce more than one mRNA that usually encode for different ORFs by a process called differential or alternative splicing. This process relies on the existence of alternative 5' and/or 3' splice sites and/or the repression of splice sites, two processes that are mediated by RNA-binding proteins (RBPs), and influenced by *cis*-regulatory sequences called ESEs (exonic splicing enhancers) and ESSs (exonic splicing silencers). Alternative splicing is immensely prevalent in mammals. In humans, for example, 95% of multiexon genes are alternatively spliced (Pan et al. 2008). This figure is similar in other mammals, with the mammalian clade average being 80% (Chen et al. 2014). Interestingly, there is a strong positive correlation between the rates of alternative splicing and organismal complexity in major animal groups, as measured by cell-type

number (Chen et al. 2014).

Importantly, the final splicing outcome of a pre-mRNA can be influenced by chromatin structure and histone modifications (Luco et al. 2011; Zhou et al. 2011), as well as transcriptional routines (Kornblihtt et al. 2004). In an illustrative example of the integration between alternative splicing and other levels of gene regulation, RBPs of the Hu family (like HuR) mediate the regulation of alternative splicing by binding to a pre-mRNA in target sequence in mouse embryonic stem cells (see Chapter 5) and concurrent hyper-acetylation of histones in the transcribed DNA region (Zhou et al. 2011). The latter effect leads to an elevated rate of transcriptional elongation and concomitant skipping of exons (SE) in the pre-mRNAs of a reporter NF1 gene (Zhou et al. 2011).

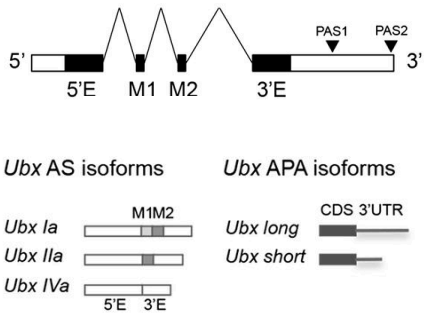
Importantly, *Hox* genes undergo alternative splicing (AS) in a number of cases. The *Drosophila melanogaster Hox* gene *Ubx* produces 6 alternatively spliced mRNA isoforms (**Figure 1.8A-B**). The outcomes of this alternative RNA processing are conserved in other *Drosophila* species like *Drosophila virilis* (Bomze & López 1994), species that diverged 60 million years ago. Alternative *Ubx* isoforms have similar 5' and 3' protein-coding exons, but may differ in three smaller sequences that lie in between the two: microexons M1 and M2 and a small extension to the 3' exon – the B element. Alternative splicing of this locus introduces a combinatorial quality to *Ubx* ORFs, in which all three optional exonic elements can be absent (isoform Iva) or present (isoform Ib); alternatively, different isoforms can lack only the B element (isoform Ia), lack the M1 microexon (IIb), or include only the M2 (IIa) or B (IVb) sequences (Bomze & López 1994). Alternative *Ubx* isoforms display a characteristic developmental pattern, indicating that the regulation of alternative splicing can be tissue-specific (Bomze & López 1994). These isoforms can also have distinct function.

For instance, the heat-shock driven expression of the *Ubx-Ia* isoform in the peripheral nervous system (PNS), but not *Ubx-IVa*, leads to a homeotic transformation at the cellular level, where the thoracic PNS is transformed into the likeness of abdominal PNS cells (Mann & Hogness 1990). Other *Hox* genes, like *Antp, abd-A* and *Abd-B* also display alternatively spliced isoforms in *Drosophila melanogaster*. Interestingly, *abd-A* forms an alternatively spliced form that does not encode for the YPWM motif in *Euperipatoides kanangrensis* (Janssen et al. 2014), part of the Onychophora – an arthropod sister-group. This observation indicates that the alternative splicing of *Hox* genes can introduce variation in in the presence/absence of key *Hox* protein motifs.

The correspondence between alternative spliced mRNAs and proteins with alternative domains is also observed in mammals, and at a transcriptome-wide scale. For instance, 62% of mouse transcription-factor (TF) loci were found to undergo alternative splicing; in turn, 68% of these events affected coding regions known to be important for TF function (Taneri et al. 2004), including the DNA binding domain (75% of cases). The remodelling of protein motifs by alternative splicing is also observed in the humans (Talavera et al. 2009) and, in conjunction with the fact that most mammalian genes are alternative spliced (see above), indicates that alternative splicing can remodel and expand the mammalian proteome directly, as well as indirectly through the introduction of functional differences between alternatively spliced TF isoforms, like those observed in the *Hox* gene *Ubx*, which may lead to the differential regulation of distinct sets of target sites. Interestingly, a number of Homeodomain genes have been shown to produce isoforms that do not encode for the Homeodomain.

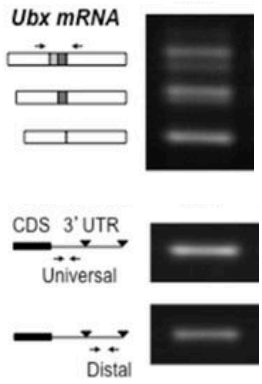
The Homeodomain genes *bicoid* and *hth* (a *Hox* co-factor, see above) have been shown to produce isoforms that lack the Homeodomain by alternative RNA splicing, during the development of *Drosophila* ((Driever & Nüsslein-Volhard 1988; Noro et al. 2006).

A



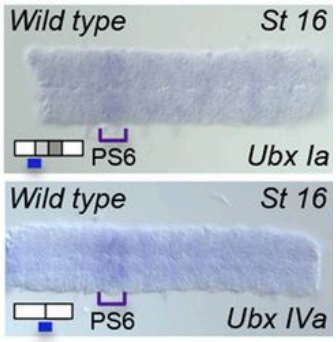
adapted from Rogulja-Ortmann *et al.* (2014)

B



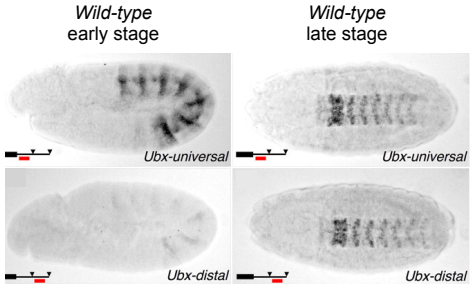
adapted from Rogulja-Ortmann *et al.* (2014)

C



Rogulja-Ortmann *et al.* (2014)

D



adapted from Thomsen *et al.* (2010)

Figure 1.8 – Ubx mRNAs undergo regulated alternative splicing and polyadenylation during Drosophila melanogaster development (legend in the following page).

Figure 1.8 – *Ubx* mRNAs undergo regulated alternative splicing and polyadenylation during *Drosophila melanogaster* development. (A) Diagram depicting the alternative mRNA isoforms that are generated in the *Drosophila melanogaster Ubx Hox* locus (adapted from (Rogulja-Ortmann et al. 2014)). Differential RNA processing of mRNAs in the *Ubx* locus generates alternative *Ubx* isoforms that differ in their open-reading frame (CDS, *Ubx AS isoforms*) and 3'UTR sequences (*Ubx APA isoforms)*. **(B-D)** Regulated expression of alternative *Ubx* mRNAs in the development of *Drosophila melanogaster* (adapted from (Rogulja-Ortmann et al. 2014) and (Thomsen et al. 2010)). **(B)** The *Ubx* locus expresses alternative mRNAs that differ in their protein-coding sequence, as well as their 3'UTRs during the development of *Drosophila melanogaster*. **(C-D)** The *Ubx* locus is expressed in the developing central nervous system (CNS) of *Drosophila melanogaster*. In this tissue, the *Ubx locus* shows expression of different mRNAs that differ in **(C)** their protein-coding sequence and **(D)** 3'UTRs. In the latter case, the alternative cleavage and polyadenylation of *Ubx* mRNAs is developmentally regulated. In early stages of embryonic development, *Ubx* exhibits short 3'UTRs; during the formation of the *Drosophila melanogaster* CNS at later stages, *Ubx* expresses mRNAs that contain longer 3'UTR sequences. These alternative 3'UTR isoforms contain distinct sets of conserved miRNA targets (Thomsen et al. 2010; Patraquim et al. 2011), and are though to mediate the differential visibility of *Ubx* mRNAs to these small RNA molecules (Thomsen et al. 2010).

In the case of Hth, the authors found that most functions of this protein can be fulfilled by the isoform that lacks the Homeodomain, including Hox-related molecular activities (Noro et al. 2006). Interestingly, *Hox* genes have been shown to produce isoforms that lack the Homeodomain in the case of the murine and human *Hoxa1* and *Hoxa9* genes, as well as the *Xenopus XlHbox2* (Fernandez & Gudas 2009). However, the regulatory mechanisms that control this alternative splicing reaction, its evolution and putative developmental roles remain largely unknown.

The *Hoxa9* gene is one of the most studied *Hox* loci of mammals (Popovic et al. 2008). *Hoxa9* is expressed in the developing forelimb, as well as the axial skeleton of mammals (see above). Additionally, *Hoxa9* is highly expressed in normal hematopoietic stem cells, and absent during their differentiation (Stadler et al. 2014). The mis-regulated expression of this gene is observed in a number of acute myeloid leukaemia (AML) patients, and correlates with poor prognoses of disease progression (Golub et al. 1999; Stadler et al. 2014). The overexpression of *Hoxa9* leads to stem-cell expansion and AML in mice, being thus sufficient to induce leukemogenicity in these tissues (Thorsteinsdottir et al. 2001; Stadler et al. 2014). Recently, a *Hoxa9* isoform that lacks the Homeodomain has been shown to be sufficient to recapitulate the leukaemogenic effects of the locus (Stadler et al. 2014). This disease phenotype is thus directly linked to the control of alternative splicing of *Hox* genes (Stadler et al. 2014).

The Homeodomain-lacking (Homeodomain-less) isoform of *Hoxa9* is expressed in a regulated manner in the developing mouse, and is particularly abundant in the embryonic genital tract, kidney, forelimb and tail (Dintilhac et al. 2004). Additionally, the production of a Homeodomain-less *Hoxa9* splice form seems to be conserved between birds and mammals (Dintilhac et al. 2004). Together, these observations introduce the notion that alternative splicing can remodel the *Hox* proteome in a

significant way, by combinatorial inclusion/exclusion of key *Hox* functional domains. They also point to the fact that the mis-regulation of *Hox* splicing can lead to disease phenotypes, implicating this regulatory level in normal physiology.

Finally, I will address the RNA processing mechanism in which transcription is terminated in eukaryotes, leading to a precise definition of the 3'end of an mRNA. In prokaryotes, there are two ways in which transcriptional termination is achieved. The first, involving intrinsic transcription terminators, relies on the formation of a short mRNA hairpin that disrupts the interaction between the nascent RNA and the DNA-RNA polymerase complex. The second, deemed Rho-dependent transcriptional termination, relies on a *cis*-regulatory region, which recruits a *trans*-acting Rho protein (Ciampi 2006). This protein then contacts the RNA polymerase, leading to the dissociation of the mRNA (Ciampi 2006). In eukaryotes, transcriptional termination relies on a mechanism similar to the latter, called cleavage and polyadenylation.

As RNA polymerase II transcribes a locus, bypassing the translational STOP codon and progressing into the 3'UTR (3' untranslated region), two proteins termed CPSF and CstF scan the nascent RNA for the presence of a polyadenylation signal (PAS, usually a AATAAA hexanucleotide (Derti et al. 2012)). These proteins travel with the Carboxy-terminal domain of RNA Polymerase II during transcriptional elongation, transferring to the PAS upon the transcription of this hexamer into the nascent RNA (Lutz 2008). These proteins then recruit other factors onto the nascent mRNA, cleaving it downstream of the PAS and adding a string of adenines at the 3'end, successfully cleaving and polyadenylating a nascent transcript, a process which effectively stops the transcription of the locus and leads to the release of a mature RNA (mRNA). As with other mechanisms of RNA processing, cleavage and polyadenylation can be regulated to produce alternative 3'ends, in a process in which alternative PASs

exist in the same 3'UTR. The process of PAS selection and 3'UTR formation is called alternative cleavage and polyadenylation (APA).

The *Drosophila Hox* genes *Antp*, *Ubx*, *abd-A* and *Abd-B* undergo regulated APA during embryonic development (**Figure 1.8D**). In early development, these genes express mRNAs with short, constitutive, 3'UTRs. In the later development of the central nervous system (CNS), mRNAs for all 4 *Hox* genes display a lengthening of 3'UTRs (Thomsen et al. 2010). This lengthening of *Hox* 3'UTRs relies, at least partly, on the RNA-binding protein ELAV, as ELAV-null embryos displayed lower expression of the long *Ubx* 3'UTR isoform (Thomsen et al. 2010). Indeed, at least 383 transcripts were shown to display 3'UTR lengthening by APA in the developing *Drosophila* CNS (Smibert et al. 2012; Hilgers et al. 2012), while showing an extensive 3'UTR shortening in the testes. Most of the mRNAs that carry longer 3'UTRs in the CNS encode for transcription factors or RBPs (Smibert et al. 2012) and include ELAV (Hilgers et al. 2012). As with the posterior *Hox* genes, ELAV was found to mediate the 3'UTR extension of number of additional genes in the CNS (Hilgers et al. 2012). This was shown to rely on specific extension-associated promoters, and on the recruitment of ELAV by a transcriptionally paused RNA polymerase II, demonstrating that the CNS-specific choice of distal PASs relies on transcriptional initiation and elongation. In most CNS-elongated transcripts, the distal 3'UTR tracts are enriched for target-sites for microRNAs (miRNAs) and RBPs (Smibert et al. 2012).

miRNAs are short regulatory RNA molecules (≈20 nucleotides in size) that associate with proteins of the RISC, and bind to 3'UTRs of target mRNAs through Watson-Crick complementarity to elicit the repression of gene expression via the promotion of either target mRNA instability or its endonucleolytic cleavage. These molecules have been show to regulate *Hox* genes in a number of contexts. The miRNA

hsa-mir-196b interacts with *Hoxb8* in the initial stages of mammalian limb development, where *miR-196* functions as a fail-safe mechanism acting upstream of Hoxb8 and Shh to "assure the fidelity of (the) expression domains" of these target genes (Hornstein et al. 2005). In *Drosophila melanogaster*, the CNS-specific long *Ubx* 3'UTR carries strong targets for miRNAs *iab-4-5p* and *iab-4-3p* (Thomsen et al. 2010). As these miRNAs are co-expressed with *Ubx* in the posterior embryonic CNS development, the deployment of alternative 3'UTRs leads to the miRNA-mediated repression of *Ubx* expression in posterior domains, and is key to the establishment of the precise patterns of Ubx protein expression in the CNS (Thomsen et al. 2010), (**Figure 1.8D**).

The formation of alternative 3'UTRs by APA is present in at least 69.1% of known human genes (Derti et al. 2012). Interestingly, the CNS of mice and humans also displays a systematic lengthening of 3'UTRs by means of APA (Miura et al. 2013). In both mice and humans, close to 2000 genes display alternative long 3'UTRs, which contain thousands of conserved miRNA target sites (Miura et al. 2013). Conversely, in glioblastoma tumours that show reduced expression of the APA factor CFIm25, 3'UTRs are systematically shortened, a factor which is linked to increased tumorigenesis (Masamha et al. 2014). Mayr and Bartel also found that cancer cells display systematic shortening of 3'UTRs (Mayr & Bartel 2009). In this study, the authors note that mRNAs with shorter 3'UTRs produce higher amounts of protein, and that in the case of oncogene IGF2BP1/IMP-1, shorter isoforms lead to a higher oncogenic effect, implicating APA in tumorigenesis (Mayr & Bartel 2009).

In a 2008 study, human alternative splicing and polyadenylation patterns were found to vary more across tissues than between individuals, indicating that these two regulatory levels can be broadly regulated in a tissue-specific manner (Wang et al.

2008). Furthermore, patterns of AS and APA were correlated across tissues, suggesting that these two RNA processing mechanisms are coordinated during the formation of human mRNAs.

As such, the mechanisms of transcriptional initiation, splicing and cleavage and polyadenylation together form an integrated level of gene expression that can be called RNA processing. Due to the fact that they provide regulatory alternatives and not vast novelties, *sensu stricto*, tandem transcription start-sites, alternative splicing and alternative polyadenylation, all constitute nested levels of gene regulation. Nucleotide substitutions might easily introduce novel splice sites or polyadenylation signals, as these are short sequences (dinucleotide in one case, hexanucleotide in the latter). By merely substituting a few nucleotides, DNA sequences can thus become subjected to these nested levels of RNA-level gene regulation during evolution. In regions like the vertebrate *Hox* gene clusters, which exhibit the highest amount of compaction even when compared with *Hox* clusters of *Drosophila* and *Amphioxus* (Duboule 2007), the introduction of novel sequences like transposons or gene translocations might highly disrupt pre-existing *cis*-regulatory information; in these case, nestedness might be one of the few kinds of regulation that is at once possible in terms of genomic context, potentially powerful in outcome, and introducing a minimal amount of mutations in pre-existing genetic loci. *Hox* genes are simultaneously robust, having a key conserved function in the patterning of the main axis of animals, and evolvable, as variations in the *Hox* code follow major morphological variations across taxa.

In the following Chapters, I study the differential RNA processing of mammalian *Hox* genes, exploring the patterns, functional consequences and evolution of this regulatory level in the *Hox* gene family, and develop the argument that the simultaneous robustness and evolvability of *Hox* genes might lie, in part, on the

acquisition of alternative RNA processing during evolution.

1.7 – Aims and outcomes of this thesis.

In the previous sections of this Chapter I argue that the *Hox* genes of mammals are subject to several levels of regulation during mammalian development, which impact both chromatin and transcriptional states, mediating both repression and activation of *Hox* gene expression across the mammalian *Hox* clusters. This in turn impacts the morphogenetic activity of these developmentally important loci. Comparably, I argue, the expression of *Hox* genes in the arthropod *Drosophila melanogaster* has also been shown to include regulated chromatin and transcriptional inputs. As with mammals, I also observe that the regulation of *Hox* gene expression by chromatin and transcription has an impact on *Hox*-controlled developmental programs.

I also argue that in *Drosophila melanogaster*, there are subsequent levels of *Hox* gene regulation that involve both the differential processing of *Hox* mRNAs and their post-transcriptional regulation by *trans* acting factors like miRNAs and RBPs. These levels have also been shown to impact *Hox* developmental programs in this arthropod, both maintaining and refining *Hox* expression during the development of *Drosophila melanogaster*. These observations introduce the question of whether, as with chromatin and transcription-level regulation, the RNA-based regulatory programs of *Drosophila melanogaster* are also at work in the establishment, maintenance and refinement of *Hox* expression patterns during the development of mammals. Moreover, they raise the more general question of whether differential RNA processing and post-transcriptional regulation have the potential to significantly impact the well-established developmental programs under *Hox* control in the mammalian clade.

I address the aforementioned questions in this thesis. In Chapter 3, I use freely available *Hox* mRNA sequences to investigate the incidence, rate and evolutionary patterns of differential RNA processing in mammalian *Hox* genes. I show that *Hox* differential RNA processing is widespread in the *Hox* clusters of mammals and shows a relationship with gene duplication in the mammalian clade; I also observe that this relationship is conserved across vertebrates. I then study the manner in which alternative mRNAs are produced in the mammalian *Hox* clusters, to show that differential RNA processing involves the coordination of multiple levels of RNA processing, which work to produce alternative *Hox* mRNAs in at least two distinct modes; I also see that paralogous *Hox* generally share differential RNA processing modes in both *Mus musculus* and *Homo sapiens*. Finally, I look at the regulatory consequences of *Hox* differential RNA processing, focusing on the formation of alternative 3'UTRs in the context of miRNA-mediated *Hox* regulation. I show that there is a segregation of miRNA targets across alternative *Hox* 3'UTR isoforms, with distal 3'UTRs displaying more numerous stronger and evolutionary labile miRNA targets, when compared to constitutive 3'UTR tracts of mammalian *Hox* genes.

In Chapter 4, I investigate the impact of *Hox* differential RNA processing on Hox protein-sequences. I use an unbiased approach to study the impact of differential RNA processing on the inclusion and exclusion of key Hox functional domains, like the hexapeptide and the Homeodomain, and report that the observed *Hox* mRNA repertoire of mammals has the potential to significantly impact the molecular function of Hox transcription factors. I then focus on the production of *Hox* mRNAs that do not encode for the Homeodomain, studying the human *Hoxa9* locus in a cell-culture experimental setup, and show that longer Homeodomain-encoding *Hoxa9* mRNA isoforms contain all *cis*-regulatory sequences for their differential RNA processing into Homeodomain-

less isoforms. Further, I show that this process occurs by a quick switch in RNA processing, which is transcriptionally dependent. Based on these results, I design experiments to assess the incidence of this mechanism *in vivo*, and report the results by others, which show that the production of differential mRNA isoforms that do not encode for the Homeodomain is regulated during the development and adulthood of *Mus musculus*, respectively in time and space. Finally, I show that the production of mRNA isoforms that do not encode for a DNA-binding domain is not restricted to *Hox* genes, being observed in other Homeodomain-carrying loci of mammals, as well as in all other transcription-factor encoding gene families. I extend these observations to show that there is an extreme conservation of this process across bilateral animals, with homologous loci displaying the conserved ability to produce mRNAs that do not encode for DNA-binding transcription factor domains across arthropods, annelids and chordates.

In Chapter 5, I use an unbiased computational approach to ask whether the 3'UTRs of *Hox* genes contain information that impacts the gene expression of host mRNAs. I focus in the context of the developing forelimb of mammals, in which the mRNA expression patterns of *Hox* genes is both dynamic and well understood, and show that the 3'UTRs of different *Hox* genes share conserved sequence motifs in direct proportion to their levels of co-expression in the early developing forelimb. I also show that this pattern does not reflect the evolutionary history of mammalian *Hox* clusters, being the likely result of convergent evolution to specific molecular environments within a complex and dynamic developing tissue. Finally, I extend these observations to the 3'UTRs of *Hox* and other genes in the developing hindbrain of *Mus musculus*, showing that in this context, as with the forelimb bud, the 3'UTRs of co-expressed genes contain shared sequence motifs.

Altogether, this work shows that the differential RNA processing of mammalian *Hox* mRNAs is widespread and relates to the evolution of *Hox* clusters in multiple ways; this regulatory level has the ability to strongly impact the molecular function of Hox proteins during the development of mammals. Finally, I propose that the 3'UTRs of mammalian *Hox* genes contain *cis*-regulatory motifs that relate to the dynamic *Hox* expression patterns during the morphogenesis of both primary and secondary mammalian axes. The work presented in this thesis suggests that further studies on the molecular control of mammalian development should take the RNA-based regulation of *Hox* gene expression into account.

Chapter II

Materials and Methods

In this chapter, I describe the materials and methods used for the elaboration of this thesis.

Bioinformatic analyses

2.1 – Batch sequence retrieval from the online database *Ensembl*.

All batch sequence downloads were performed using the *BioMart* tool available on the *Ensembl* online database (http://www.ensembl.org/biomart). The following genome assemblies were used: BDGP5 (Drosophila melanogaster), WBcel235 (*Caenorhabditis elegans*), Zv9 (*Danio rerio*), GRCm38.p3 (*Mus musculus*) and GRCh38 (*Homo sapiens*). In the cases of the latter two datasets, only the isoforms with GENCODE basic annotations were used in further analyses. The previously published protein sequences of *Branchiostoma lanceolatum* Hox genes Hox1-14, as well as Hox10 homologues in *Xenopus tropicalis*, *Ciona intestinalis and Oikopleura dioica* were retrieved from the UniProt database (http://www.uniprot.org/uniprot/Q9NAZ0)

2.2 – Estimation of protein divergence rates within Hox paralogue groups.

To estimate the protein divergence rates within each of the 13 mammalian paraloguous-groups (PGs), the protein sequences for all reference Hox protein isoforms in each PG of both *Mus musculus* and *Homo sapiens* were aligned to the corresponding *Branchistoma lanceolatum* single Hox orthologue, using the MAFFT algorithm (Katoh & Standley 2013), see section **2.1**. These alignments were used to construct a phylogeny tree for each PG, using the Neighbour-Joining method with a JTT substitution model in

the MAFFT algorithm (Katoh & Standley 2013). For each PG, the average protein divergence was calculated by adding all the individual tree-branch lengths between each terminal paralogue and PG ancestral node, and dividing this value by the total number of paralogues in each group.

2.3 – Categorization of mammalian Hox differential RNA processing events for individual protein-coding isoforms.

For each of the 39 *Hox* loci in both *Mus musculus* and *Homo sapiens*, all the *Hox* RNA isoforms with strong experimental support were downloaded from *Ensembl BioMart* (see section **2.1**) and aligned to each other using either the online MAFFT algorithm, the *Ensembl Transcript* comparison tab or the *Serial Cloner* software (http://serialbasics.free.fr/Serial_Cloner.html). A reference isoform was selected for each locus (the longest isoform encoding for a homeodomain), and used to pinpoint individual differences in each of the alternative isoforms. The resulting alignments were scanned for the existence of variant splice-sites, START or STOP codons and polyadenylation sites. The differential mRNA processing events were then categorized using the alternative transcript event framework in (Wang et al. 2008), to which the annotation of tandem Transcription Start Sites (tTSS) was added. To study correlated differential RNA processing events, I employed the *PerformanceAnalytics R* package (http://cran.r-project.org/web/packages/PerformanceAnalytics/index.html).

2.4 – Hierarchical-clustering analyses.

All data was loaded into R in *.csv* format and transformed into a data matrix.

This data matrix was then transformed into a Euclidean distance matrix using the R function *dist()* (https://stat.ethz.ch/R-manual/R-patched/library/stats/html/dist.html) and hierarchically clustered using the Hierarchical Clustering R function *hclust()* (https://stat.ethz.ch/R-manual/R-patched/library/stats/html/hclust.html) and the *average* (UPGMA) agglomeration method. Rows and Columns were independently clustered, and the results were used to in conjunction to construct a Heat Map, using the *Enhancead Heat Map heatmap.2()* function contained in the *gplots* R package (http://cran.r-project.org/web/packages/gplots/). For bootstrap measurements, I used the AU measurement in the *pvclust* R package (R. Suzuki & Shimodaira 2006).

2.5 – miRNA targeting predictions in the context of alternative Hox 3'UTR formation.

All experimentally validated *Hox*-miRNA interactions were downloaded from miRTarBase (Hsu et al. 2014). *De novo* miRNA targeting predictions for alternative *Hox* 3'UTR sequences were performed using the PITA miRNA targeting prediction tool (Kertesz et al. 2007). In the latter analyses, only strongly predicted miRNA targets were used, corresponding to a predicted $\Delta\Delta G \leq -10$ as per the recommendation of the authors. Alternative *Hox* 3'UTR isoforms were defined according to GENCODE-annotated 3'-ends (Harrow et al. 2012).

2.6 – Pre-computed protein-domain predictions.

Batch protein-domain predictions were downloaded from *BioMart* (see **2.1**), using the pre-computed *SMART ID*, *PROFILE ID*, *PRINTS ID*, *PFAM ID* and

TIGRFam ID predictions in the *Protein domains and families/Domains* section of *BioMart/Attributes/Features*. The pre-computed *Interpro ID*s were also used (*Protein domains and families/Interpro)*. In cases where the protein-domain predictions were inexistent or of insufficient quality, individual amino-acid sequences were submitted to the InterProScan tool (http://www.ebi.ac.uk/interpro/interproscan.html) and the results checked manually. For the bioinformatic query on the existence of Homeodomain-less RNA isoforms of annotated Homeodomain genes, the *SMART ID* SM00389 (http://www.ebi.ac.uk/interpro/entry/IPR001356) was first used as a filter (*BioMart/Filters/Protein domains and families*). The resulting gene list was then used as the only filter (*BioMart/Filters/Gene/Ensembl Gene ID)* to submit another *BioMart* query that returned a list of transcript IDs for each gene, regardless of domain predictions. This list of transcript IDs was then submitted as a filter (*BioMart/Filters/Gene/Ensembl Transcript ID)*, along with batch protein-domain predictions as described above. For each isoform, the absence of Homeodomain predictions was ascertained and confirmed independently using InterProScan (see above, this section). The same method was used for leucine zipper, zinc finger and helix-loop-helix DNA-binding Domain predictions (SMART IDs SM00338, SM00355 and SM00353, respectively).

2.7 – Unbiased Hox protein-motif predictions.

To predict Hox protein domains in an unbiased manner, I submitted the translations of all GENCODE-annotated mRNA isoforms of *Mus musculus* and *Homo sapiens* (see section **2.1**) to the MEME motif-search tool (Bailey et al. 2009). For this analysis, I used the Normal Mode setting of the MEME software, querying Hox

proteins for a maximum of 30 ungapped motifs of 6-50 amino acids in size. I then annotated the presence and absence of each motif in each Hox protein in a Microsoft Excel table, respectively using the numerical values "1" and "0". This table was then used in subsequent hierarchical clustering analyses (see section **2.4**).

2.8 – Computational representation of Hox spatial expression patterns in the developing forelimb of Mus musculus.

First, I retrieved published images of RNA *in-situ* hybridizations for *Hoxd11* at 11.5 days *post coitum* (d.p.c.) in the forelimb (Bruneau et al. 2001) and digitally saturated them, so that the observed expression patterns became binary (absence or presence). I then superimposed a square matrix with 450 (25X18) equally-distributed points on this image, and assigned a numerical value to each of the points: "1" - if a specific point overlapped totally or partially with an area showing *Hoxd11* expression, and "zero" - "0" – if a point fell on a region with no *Hoxd11* expression. I obtained a square matrix with "1" and "0" values that was linearised, by concatenating all rows in tandem, in a top-to-bottom order, rendering a 672-point line with binary values. Given the difficulty in finding high-resolution records of *Hox* RNA *in-situ* hybridizations for all 20 *Hox* genes that are expressed in the forelimb, I compared the resulting *Hoxd11* binary expression profile with a profile generated by the same method described above, and using a detailed *Hoxd11* expression model published by *Zakany & Duboule* (Zakany & Duboule 2007). The main aspects of 11.5 d.p.c. forelimb *Hoxd11* expression (e.g. mostly distal, split into two smaller domains) were recapitulated by the late phase model. As such, I decided to use similar expression models for all 20 forelimb-expressed *Hox* genes, in both early and late waves of forelimb expression. The late-

stage binary expression profiles were performed as described above. The early-stage profiles used a slightly smaller initial matrix with 350 points (25X14), as the forelimb is smaller at this stage; all subsequent analyses were performed in the same manner as with the late-stage forelimb expression models. The 20 profiles were then put together in a matrix, with each row representing the presence or absence of expression of a given *Hox* gene in 800 forelimb spatiotemporal data points. This matrix was submitted to a Hierarchical Clustering analysis (please see section **2.4**).

2.9 – Bioinformatic search for rhombomere-specific gene-expression in developing hindbrain of Mus musculus.

I used the *Gene Expression Data Query* tool in the GXD (Gene eXpression Database; http://www.informatics.jax.org/gxd) to retrieve all genes reported, by RNA *in-situ* hybridization, to be expressed in each of the 8 mouse hindbrain rhombomeres at 8.5-10.5 d.p.c. This time-window was chosen to coincide with the time at which rhombomeres form (9 d.p.c.) during mouse hindbrain development. This list of genes was then divided in two sets. The first included 32 genes that have segmentally-restricted expression ("rhombomere-restricted"). The 3'UTRs of these genes were used as *positive sequences* in the MEME motif analysis (please see section **2.11**). The remaining genes, those whose expression transgressed rhombomere boundaries during this time-window, being thus detected in 2 or more rhombomeres, were ascribed to the second set of sequences, the *negative sequences* set (please see section **2.11**). For the rhombomere-restricted genes, I created a binary matrix, with genes as rows and rhombomeres as columns, by ascribing a numerical value of "1" or "0" to the presence or absence of expression, respectively, of a given gene in each of the 8 rhombomeres.

This matrix was then submitted to hierarchical clustering (please see section **2.4**) to construct a cladogram reflecting expression similarities between the 32 genes analysed.

2.10 – Computational representation of spatial gene expression patterns in the *C. elegans* germline.

The germlime expression patterns of 30 *C.elegans* genes have been reported in (Merritt et al. 2008). The authors cloned the 3'UTRs of 30 *C. elegans* genes (normally expressed in the germline) downstream of a green fluorescent protein-histone H2B fusion (GFP-H2B), and drove expression of each construct in all germ cell types using the *pie-1* promoter. The results showed that for 25 of these genes, the 3'UTRs were sufficient to spatially restrict the expression of each construct in a manner resembling the host gene's endogenous protein-expression domain. I thus constructed a matrix representing the expression of *C. elegans* genes in the germline, with each of the 25 genes representing a row, each of the 8 spatial domains analysed in the aforementioned study as columns, and ascribing values of "1" and "0" to the presence or absence, respectively, of endogenous protein expression for each gene in each spatial coordinate. This matrix was then submitted to hierarchical clustering (please see section **2.4**).

2.11 – Computational search for 3'UTR-enriched motifs.

3'UTR sequences were submitted to the MEME motif-search tool ((Bailey et al. 2009), see section **2.7** for a similar query using Hox protein isoforms). A discriminative motif discovery analysis was performed on the given strand only, looking for a maximum of 30 motifs $(6\text{-}10)_{mer}$-long. As such, the algorithm was prompted to find

motifs that were both present in the set of sequences provided (the *positive sequences* set) and absent in the *negative sequences* set. In the case of the mouse forelimb analysis, as a *positive* set I submitted sequences for the longest annotated 3'UTR for the 20 *Mus musculus HoxA/D* genes, as well as the 3'UTRs of the respective *Homo sapiens* orthologues in order to enrich our analyses in ultra-conserved mammalian motifs. As the *negative sequences*, I submitted the longest 3'UTR sequences pertaining to all 19 *HoxB/C* genes of both *Mus musculus* and *Homo sapiens*. These genes have minimal phenotypic effects on limb morphogenesis when mutated (Zakany & Duboule 2007). In the case of the mouse hindbrain analysis, I submitted the longest annotated 3'UTRs of rhombomere-restricted *Mus musculus* genes (please see section **2.9**) as the *positive sequences*, and the longest annotated 3'UTRs of genes whose expression that transgresses rhombomere boundaries as our *negative sequences* set. In the case of the *Caenorhabditis elegans* germline analysis, the longest annotated 3'UTRs of the 25 genes experimentally shown to spatially restrict the expression of a reporter gene in the *C*. elegans germline were used as a positive set. No sequences were used as the *negative sequences* set in this case. In all three cases, the data was transformed into a matrix, with genes represented as rows, each of the 30 motifs represented as columns, and the presence or absence of a given motif represented as the numerical values "1" and "0", respectively. Each matrix was then hierarchically clustered (please see section **2.4**), and the results were compared with the hierarchical clustering results of the respective gene expression patterns (please see the following section).

2.12 – Matching 3'UTR motifs to gene expression patterns using the Subtree pruning and regrafting (SPR) algorithm.

To compare the results of hierarchically clustering genes base on their 3'UTR-motif similarities and their spatial expression pattern similarities, I employed a heuristic algorithm, *the Subtree pruning and regrafting (SPR)* algorithm (Goloboff 2008; Goloboff et al. 2008). SPR is commonly used in phylogenetic analyses to determine optimal tree structure, as this method compares trees that are composed of the same clades but show different topologies. In order to do so, the algorithm determines the minimal number of "pruning" and "regrafting" (cut and paste) operations (or *moves*) that the tree branches of tree A have to undergo in order to arrive at tree B. The least the number of operations, the more similar trees A and B are. I used a version of the Tree Analysis using New Technology (TNT) software (Goloboff et al. 2008), modified by our collaborator Martín Ramirez at the *Museo Argentino de Ciencias Naturales* (MACN) to run SPR analyses, as well as to perform the appropriate statistical validations (see below, this section). For each experiment, the tree topology of a cladogram resulting from hierarchical clustering (please see section **2.4**) was reconstructed manually in the TNT command line. This was performed for genes organised according to their 3'UTR motif information (please see the previous section) – Tree A – and for the same genes, now organised according to their expression patterns – Tree B (please see sections **2.8**, **2.9** and **2.10**). These trees were then compared in TNT, using the SPR method, and a number of SPR *moves* (operations) was returned. I then separately randomised each of the two trees 10.000 times while keeping the other constant, and performed and SPR analysis on each random tree-real tree pair (20.000 pairs in total). I asked how many times a random tree *A'* is as successful as our original tree A, that is, needing as few SPR moves between tree *A'* and tree B as our original tree. This provided a measure of statistical significance to the matching between 3'UTR information and gene expression.

Experimental analyses

2.13 – Minipreparation of plasmid DNA.

A copy of the *Hoxa9* human cDNA Clone (*SC321224*; untagged) was purchased from *OriGene* (http://www.origene.com/human_cdna/NM_152739/SC321224/HOXA9.aspx). This plasmid contains a Citomegalovirus (CMV) promoter fused to the cDNA of the human *Hoxa9-001* isoform (ENST00000343483) and will be refered hereafter as the "pCMV-*Hoxa9*" plasmid. The pCMV-*Hoxa9* plasmid was used to transform competent *E.coli* bacteria. Plasmid DNA was isolated from bacterial cultures using the *QIAprep Spin Miniprep Kit* (*QIAGEN*) following the instructions provided by the manufacturer. A 5 mL volume of overnight culture were spun to pellet the cells (3 min, 3000 r.p.m.) and the supernatant was discarded. Confirmation of the recovered plasmid identity was done via PCR using plasmid-specific primers. The confirmation of the recovered plasmid yield was done using a *Picodrop* spectrometer.

2.14 – Cell culture techniques.

HEK293-EBNA cells were obtained from a running culture in Guy Richardson's Laboratory (http://www.sussex.ac.uk/profiles/2231). Cell-cultures were kept in an incubator at 37ºC with 5% CO_2. T75 flasks with 25 mL of Dulbecco's Modified Eagle Medium (DMEM) were used to culture the cells. The medium was supplemented with 10% Fetal Bovine Serum (FBS), 1% Penicillin-Streptomycin (PS) and 1% L-Glutamine

(L-Glu). The cultures were regularly passaged after reaching 70% confluency, which occurred 1.5 times a week. All passages were performed in a 70% ethanol-sterilised laminar flow hood, and consisted in the removal of culture medium, followed by a washing step with sterilised PBS 1X (to remove any remaining medium), and the addition of 2 mL of Trypsin/EDTA solution (TE). The culture was then placed at 37ºC for 2 minutes, to allow for the cells to dissociate from the flask and each other. An 8 mL volume of fresh medium was then added to the 2 mL of Trypsin/EDTA/Cellular solution, as the FBS present in the fresh medium blocks any remaining activity of Trypsin. 1-10% of this solution was then used to passage the cells to a fresh T75 flask with fresh supplemented DMEM.

2.15 – HEK293-EBNA transfections with plasmid pCMV-Hoxa9.

For transfections, trypsinised cells were seeded in 6-well plates at a low concentration (1-3% of the trypsinised cell solution - see previous section). These cells were then transfected at 80% confluency with 1.5 µL of pCMV-*Hoxa9* plasmid using Lipofectamine 3000 (Invitrogen) as described by the manufacturer. In order to confirm the efficiency of transfections, I performed co-transfections of the pCMV-*Hoxa9* plasmid with the *pmaxFP-Green-N* vector (http://www.addgene.org/browse/sequence_vdb/3525/), and visualised GFP expression in an inverted fluorescence microscope. The *pmaxFP-Green-N* vector was a generous donation from our colleague Dr. Christopher Sampson at the Juan Pablo Couso Laboratory in the University of Sussex.

2.16 – Blocking transcriptional activity in HEK293 cells.

In order to block transcriptional activity in cells, actinomycin D (5 μg/mL, Sigma) was added to 6-well plates containing HEK293-EBNA cells at 80% confluency, three hours after transfection with pCMV-*Hoxa9*. The medium was then kept unchanged for 16 hours, when RNA extractions were performed (see next section).

2.17 – RNA extraction.

RNA was extracted from cells by adding 1000 μl of TRI Reagent (Sigma) to each of the wells in a 6-well plate, following the manufacturer's protocol. The cells were then re-suspended and homogenised in TRI reagent by pipetting. After homogenization, the cellular homogenate in TRI reagent was placed in 1.5 mL microcentrifuge Eppendorf tubes, and incubated for five minutes at room temperature to dissociate nucleoprotein complexes. RNA was separated from DNA and proteins by adding 200 μl of RNase free Chloroform, mixing and incubating for fifteen minutes at room temperature. The different phases – aqueous phase (RNA), interphase and organic phase (DNA and proteins) – were separated by 15 minutes of centrifugation at maximum speed at 4ºC, and the aqueous phase (colourless top layer) was transferred to a new tube. RNA was precipitated with 500 μl of Isopropanol at -80ºC for 1 hour to overnight, followed by centrifugation at maximum speed for half an hour at 4ºC. Precipitated RNA was washed in RNase free 75% ethanol, resuspended in nuclease-free water and stored at -80ºC. RNA concentration was measured in a Picodrop spectrometer. All steps were performed in RNase-free conditions.

2.18 – cDNA synthesis.

After DNAse I (NEB) treatment (as per the manufacturer's protocol), cDNA was synthesised from 1µg aliquots of total RNA using oligo(dT) primers (2 µL from a 50 µM stock solution) and the RETROscript kit (Ambion), by following the protocol for the 'Two-step RT-PCR with heat denaturation of RNA' procedure provided by the manufacturer. 1 µg of total RNA was combined with oligo (dT) primers and nuclease-free water, then denatured at 85ºC before the addition of the remaining RT reagents: 10X RT buffer, dNTPs (2 µL form a stock solution containing 2.5 mM of each dNTP), RNase inhibitor (0.25 units), and the Mu-MLV Reverse Transcriptase (2.5 units). Reverse transcription of cDNA was done at 42ºC for 1 hour, followed by 50ºC for 30 minutes. The reaction was stopped by inactivating the reverse transcriptase at 92ºC for 10 minutes. For each sample, a 500 ng of DNAse-treated RNA was mixed with 10 µL of nuclease-free water and reserved at -20ºC, to be used as no-RT controls. Newly synthesised cDNA was stored at -20ºC until ready to use in PCR reactions.

2.19 – Polymerase Chain Reactions (PCRs).

PCR reactions were prepared on ice to a final volume of 25 µl as follows: 2.5 µl of 10x PCR Buffer (New England Biolabs), 0.5 µl of 10 mM dNTP mix (New England Biolabs), 1 µl of each forward/reverse primer (10mM each, see **Table 2.1**), 0.25 µl of standard Taq DNA polymerase (New England Biolabs), 1 µl of cDNA and 18.75 µl of nuclease-free water. PCR was performed using *a BioRad T100™ Thermal Cycler* PCR machine with the following conditions:

1 cycle: *extended DNA denaturation at 95ºC for 5 minutes*

33 cycles: *template denaturation at 95ºC for 30 seconds*

primer annealing at 55ºC-65ºC for 45 seconds

extension at 72ºC for 30 seconds

1 cycle: *final extension step at 72ºC for 7 minutes*

hold: *4ºC*

All primers were optimised by conducting a gradient PCR using genomic DNA as a template, in conditions identical to the aforementioned, with an annealing temperature varying between 55ºC and 65ºC. This confirmed an optimal *endogenous Hoxa9* primer-annealing temperature of 55ºC. The optimal annealing temperature for the *GAPDH* and *c-myc* primer-pairs was determined to be 59ºC. For the *Hoxa9*-plasmid primers, the optimal annealing temperature was determined to be 65ºC. The primer list can be found in **Table 2.1** (below). Expression values were normalized using reference gene *GAPDH*. At least three independent biological replicates were performed. Two negative controls were always performed: (i) genomic contamination control in the RNA sample – PCR performed with RNA as a template - and (ii) a no template control.

Table 2.1 – PCR primers

RT-PCR Primers	Primer sequence (5' to 3')		Source
***Hoxa9*-endogenous**	FWD	GGGCAACTACTACGTGGACT	This study
	REV	TTGTTTTCAGAGAAGGCGCC	
GAPDH	FWD	GTCAAGGCTGAGAACGGGAA	This study
	REV	CAAAGGTGGAGGAGTGGGTG	
c-myc	FWD	GACTCGGTGCAGCCGTATTT	This study
	REV	TGTCGTTGAGAGGGTAGGGG	
***Hoxa9*-plasmid**	FWD	GGCCGGGAATTCGTCGACTGG	This study
	REV	AGGGGCACCGCTTTTTCCGA	

2.20 – Agarose gel electrophoresis.

RNA and PCR products were visualised in an agarose gel electrophoresis. Agarose gels were made at 0.8% concentration (w/v), by dissolving agarose in 100 mL of SB buffer (1x). The mixture was heated at in the microwave at 900W for 1.5 minutes, mixed, and heated again for 30 seconds until the agarose was completely homogenised, before being cooled in warm water. After cooling down, 0.4 µg/ml of ethidium bromide (EtBr) was added to the liquid agarose before pouring the mix into a gel cast. Samples were prepared in 1x loading buffer (New England Biolabs), loaded into the wells of the gel alongside a 100bp and 1Kb DNA ladder (New England Biolabs) and subjected to

electrophoresis in 1x SB Buffer. Gel pictures were taken using an *Uvidoc* gel documentation system (*Uvitec Cambridg*e) and *UviPhotoMW* image analysis software. Quantification of the gels was performed using imageJ.

2.21 – DNA Sequencing.

Agarose gel bands were cut on a UV light box, using sterilised scalpels for each band. Each agarose portion containing the DNA of interest was placed in a new 1.5 mL Eppendorf microcentrifuge tube. The weight of each agarose portion containing the dsDNA of interest as then determined using a balance, and the DNA was extracted from the agarose/EtBr-DNA mix using the QIAquick Gel Extraction Kit (*QIAGEN*) as per the instructions of the manufacturer. The resulting DNA was quantified using a Picodrop Spectrometer, and sent for sequencing. Sequencing was performed by Eurofins Genomics (https://www.eurofinsgenomics.eu).

Chapter III

The production of Hox *mRNAs by differential RNA processing in mammals*

3.1 – Chapter Overview

In this chapter, I explore the production of alternative RNAs in mammalian *Hox* genes by differential RNA processing (DRP), as well as its evolution in vertebrates. I first compile previously reported and well-supported mRNA sequences arising from *Hox* loci in *Homo sapiens* and *Mus musculus* models using the sequences of zebrafish *Hox* as an out-group. I find that *Hox* genes display low and heterogeneous rates of DRP, which are conserved across all vertebrates. I also observe an inverse correlation between the rate of protein evolution after gene-duplication and the average number of alternative mRNAs produced in mice, suggesting that the evolution of the two is linked in the mammalian lineage. I next produce a catalogue of coordinated mammalian *Hox* mRNA processing events, and show that the 3' untranslated regions of *Hox* genes are extensively remodelled, that alternative *Hox* mRNAs are mainly formed by two distinct successions of RNA processing events that involve the joint regulation of transcription, splicing and polyadenylation, and that for 10 of the 13 paralogue groups, the type of alternative mRNA processing that a given *Hox* gene undergoes is shared with its paralogues, suggesting that specific mammalian *Hox* DRP patterns arose at the base of the mammalian lineage. Next, I study the most frequent individual DRP event, tandem 3'UTR formation, and show that miRNA-*Hox* interactions are more prevalent in the distal 3'UTRs of *Hox* genes. I also use the full miRNA complement of mammals to show that i) the miRNA target complement of proximal and distal 3'UTRs is negatively correlated in most cases, ii) that constitutive 3'UTRs of homologous *Hox* genes share a number of target sites between *M. msuculus* and *H. sapiens* and iii) that miRNA targets in distal 3'UTRs of orthologous *Hox* genes show low conservation in mammals.

Together these results show that differential RNA processing significantly

remodels the *Hox* transcriptome, and indicate that this gene regulatory level has implications for the molecular regulation of *Hox* gene expression.

3.2 - Results

3.2.1 – Hox genes show evidence of alternative RNA production in mammals.

Differential RNA processing is a gene regulatory mechanism by which different mRNAs are produced from a single gene by means of alternative Transcription Start-Sites (TSS), alternative splicing of exons (AS) or alternative cleavage and polyadenylation (APA) of 3'Untranslated regions (3'UTRs). All these are mediated in part by alternative regulatory signals in *cis*, and often work in conjunction to produce alternative mRNAs that differ in their untranslated or/and protein-coding regions. This integrative level of gene regulation leads to the diversification gene expression at the RNA level, which in turn influences the rates and kinds of translated proteins, leading to the diversification of expression at the protein level. Based on the incidence and functional implication of DRP in the *Hox* genes of arthropods (see Chapter 1), I wondered whether mammalian *Hox* gene families also displayed evidence of this mechanism. In mammals, the *Hox* gene complement consists of 39 *loci* organized in four genomic clusters deemed A, B, C and D **(Figure 3.1A)**. Each of the clusters lies in a different chromosome, and is evolutionarily related to the other clusters through two rounds of genomic duplication that occurred at the base of the vertebrate phylogenetic lineage. As such, *Hox* genes within each of the clusters have paralogues (meaning homologues produced by gene duplication) in other clusters. The paralogue groups (PGs) are numbered 1 to 13, and the conjunction of *Hox* cluster placement and

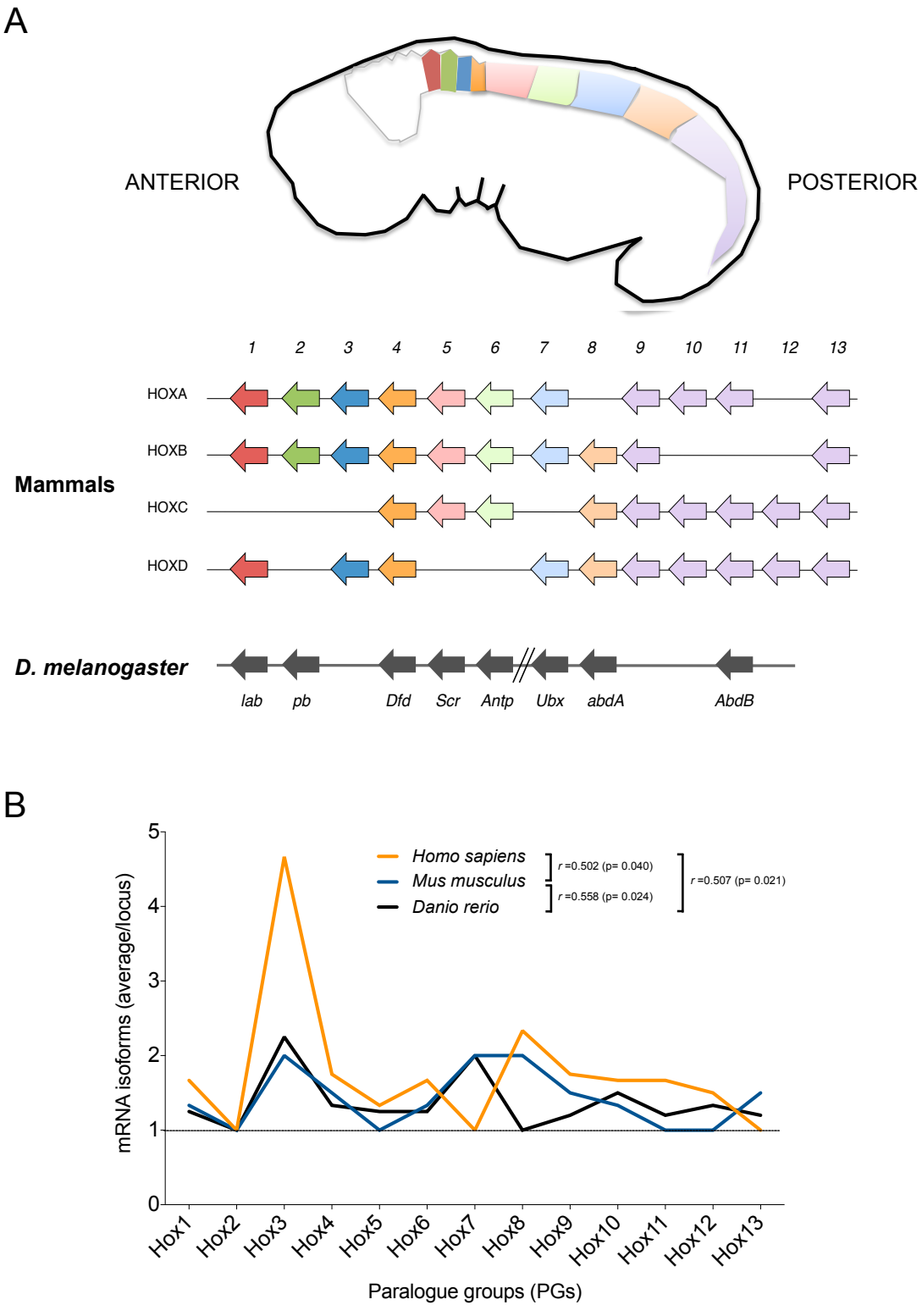


Figure 3.1 – Mammalian *Hox* genes display a conserved production of alternative mRNAs by differential RNA processing (legend in the following page)**.**

Figure 3.1 – Mammalian *Hox* genes display a conserved production of alternative mRNAs by differential RNA processing. (A) Organisation of the *Mus musculus* and *Drosophila melanogaster Hox* genes clusters. The 39 *Hox* genes of mammals are distributed across four clusters, lying in four different chromosomes. There are 13 paralogue groups (PGs), representing sets of *Hox* genes that share common ancestry by gene duplication. During mammalian development, *Hox* genes display spatial colinearity between genomic position and expression patterns along the A-P axis (*e.g. Hox* genes of PG 1 are expressed more anteriorly than *Hox* of PG13). The mammalian *Hox* clusters are homologous to the single *Hox* cluster of *Drosophila melanogaster*. In this organism, *Hox* genes are further divided into two complexes, ANT-C and BX-C, and also display spatial colinearity during their expression along the A-P axis of developing embryos. **(B)** Alternative production of *Hox* mRNAs in the Vertebrate *Hox* clusters. The average production of alternative mRNAs is heterogeneous across PGs. PGs3 and 6-9 show enrichment of *Hox* mRNA isoforms in both *Homo sapiens* and *Mus musculus* (r=0.502, p=0.040). This profile is also observed in the *Danio rerio Hox* genes, indicating that this pattern is conserved across Vertebrates (*Danio rerio-Homo sapiens:* r=0.507, p=0.021 ; *Danio rerio-Mus musculus*: r=0.558, p=0.024).

paralogue group gives each mammalian *Hox* gene a coordinate that is usually used as its identifier (e.g. *Hoxa1, Hoxb1, Hoxa3, Hoxd13* in **Figure 3.1A)**, (Scott 1993).

In order to analyse differential RNA processing (DRP) patterns occurring in mammalian *Hox* genes, I first retrieved all the previously annotated, experimentally well-supported protein-coding *Hox* RNA isoforms from *Ensembl*, using the BioMart tool (see Chapter 2 for specific criteria). This defined a conservative dataset with which to explore DRP in the 39 mammalian *Hox loci*.

Looking at the rates of *Hox* DRP, I first found that in total, the 39 *Hox* genes of mammals show the production of 56 isoforms in *M. musculus*, and 69 in *H. sapiens*. This yields an average of 1.44 isoforms per locus in the case of *M. musculus*, and 1.77 RNA isoforms in the case of *H. sapiens*. Both averages are below "2", indicating that some genes don't produce alternative isoforms. Indeed, the 56 alternative isoforms produced by *M. musculus Hox loci* stem from 13 of the 39 *Hox* genes (34%). In the case of *H. sapiens*, I find that the 69 alternative RNA isoforms retrieved are produced by only 18 of the 39 *Hox* genes (47%). Both figures stand in stark contrast to the most recent reports of DRP for mammalian *loci*. As an example, a recent, comprehensive survey of alternative splicing in *Eukarya* found that this DRP mechanism alone is present in 81.4% and 87.5% of *M. musculus* and *H.sapiens* loci, respectively (Chen et al. 2014). This indicated that our conservative dataset might be exceptionally circumscribed, leading to the under-representation of *Hox* DRP. One possibly is that this is due to the low sensitivity of the techniques employed in collecting the GENCODE data; most RNA isoforms in the GENCODE annotation are supported by low-throughput techniques like cDNA and EST sequencing, whereas most recent transcriptomic studies use RNA-seq, a high-throughput method.

I thus wondered if this impoverishment in *Hox* DRP reflected a systemic under-

representation of alternative RNA isoforms in the GENCODE annotation. To test this, I used the same database to retrieve all annotated protein-coding transcripts for both mice and humans, and compared the number of total isoforms with the number of annotated protein-coding loci in the respective genomes. I find that each of the annotated 21783 protein-coding *loci* in humans produce 2.71 protein-coding isoforms per locus on average, while the 22151 protein-coding *loci* of mice produce 1.75 protein-coding isoforms per locus on average. Furthermore, the production of 2 or more well supported RNA isoforms is observed in 14378 (66%) of GENCODE-annotated human protein-coding genes, while the figure is 10608 (48%) for *M. musculus* genes. Compared with recent studies that indicate the production of alternatively spliced transcripts in 95% of human multiexon genes (Pan et al. 2008), and 92–94% of all human genes (Wang et al. 2008), I suggest that, perhaps unsurprisingly due to the conservative nature of the dataset, DRP events are indeed underrepresented in the GENCODE annotation. However, when I define the overall GENCODE DRP rates as 100%, and normalize the percentage of DRP in *Hox* genes in relation to this value, I find that *Hox* DRP is still lower than the expected in mammals, being 57% in mice and 62% in humans. The production of alternative RNA isoforms by *Hox* genes is thus below the average, even if I take into account our concerns about the GENCODE annotation coverage.

The *Hox* genes of mammals are notoriously redundant at the functional level. For example, Wellik and Capecchi (Wellik & Capecchi 2003) found that for both *Hox10* and *Hox11* paralogue groups - which include 3 *Hox* genes each - the conspicuous effects of these genes on *Mus musculus* skeletal morphology only become apparent after 5 of the 6 alleles are mutated. As such, one would expect the different protein-coding isoforms of a given paralogue group to functionally complement each other. According to this functional sharing hypothesis, the transcriptomic diversity of

one *Hox* gene should be analysed in the context of the evolutionary history of the gene, with special regard to its duplication history, which is to say, in the context of the transcriptomic diversity of other genes in the same PG.

To explore this idea, I re-analysed our data by first binning all annotated *Hox* isoforms into each of the 13 PGs, and then dividing the total number of isoforms per PG by the number of paralogues in each group. I find that, in the context of gene duplication, 10 of the 13 PGs (79% of *Hox* genes) produce, on average, more than one isoform per locus, while 9 of the 13 PGs produce possess DRP in the case of *M. musculus*, representing 74% of *Hox* genes **(Figure 3.1B)**. Moreover, the rates of DRP for each PG are conserved across mammals (Pearson r=0.502; p=0.040), being specially augmented in the genes of PG3 (2 isoforms per locus in the case of mice, 4.6 in the case of humans), and PGs 6-9, indicating that the production of alternative protein-coding RNAs is specially important in genes of these groups. I chose to include in our analysis the well-supported alternative *Hox* isoforms in the most recent zebrafish genome release (Zv9), and found that the same profile of mammalian DRP rates is observed across PGs in this organism (zebrafish-human: Pearson r=0.507; p=0.021. zebrafish-mouse: Pearson r=0.558; p=0.024). Taken together, these results indicate that the rates of differential RNA processing vary across *Hox* duplication groups (PGs), and that this pattern is conserved across vertebrates.

The fact that *Hox* PGs display differential rates of RNA production, and that these are conserved beyond the mammalian lineage, being conserved across vertebrates, suggests that the history of *Hox* gene evolution provides an appropriate context for study of DRP in vertebrate *Hox* genes. These observations led us to inquire whether the incidence of DRP might relate to the evolutionary history of *Hox* paralogues themselves. To this end, I first estimated the average protein divergence within each PG

by retrieving the longest homeodomain-containing protein isoform for each of the 39 *Hox loci* in both mice and humans, as well as the longest known protein for *Hox genes Hox1- Hox13* in *Branchiostoma lanceolatum* (a cephalochordate). The latter sequences provided an out-group to study the evolutionary rates each of the PGs. I then performed an alignment followed by a Neighbour-Joining phylogeny for each PG using MAFFT (JTT substitution model, see Chapter 2) (**Figure 3.2A**). This allowed us to estimate the average number of amino acid substitutions per gene per PG since the first *Hox* cluster duplication at the base of vertebrates. I then compared this figure with the aforementioned average production of alternative *Hox* RNAs per PG (**Figure 3.2B**). I find that in the case of most mammalian *Hox* PGs, no significant correlation exists between protein evolution rates and the average amount of isoforms produced for each PG (**Figure 3.2B)**. In the case of PGs 9-13 however, I find a statistically significant negative correlation between within-PG protein divergence rates and the average amount of isoforms produced by *Hox* of the same PGs (Pearson r = -0.98, p=0.004) (**Figure 3.2C**). This result suggests that there is a relationship between the amount of isoforms produced by posterior *Hox* genes, and the evolution of the *Hox* protein-coding loci themselves. In light of these results, I hypothesize that in posterior *Hox* genes, rates of DRP are constrained by the evolution of the protein-sequences themselves, being lower for fast-evolving genes. In other PGs, which underwent slower rates of protein divergence in the vertebrate clade, relaxed selective pressures led to the accumulation of DRP in these loci (**Figure 3.2D**). In both cases, a diverse expression output exists which can decrease the genetic load that arises from redundant protein production. This may lead to decreased evolutionary pressures for *Hox* locus divergence

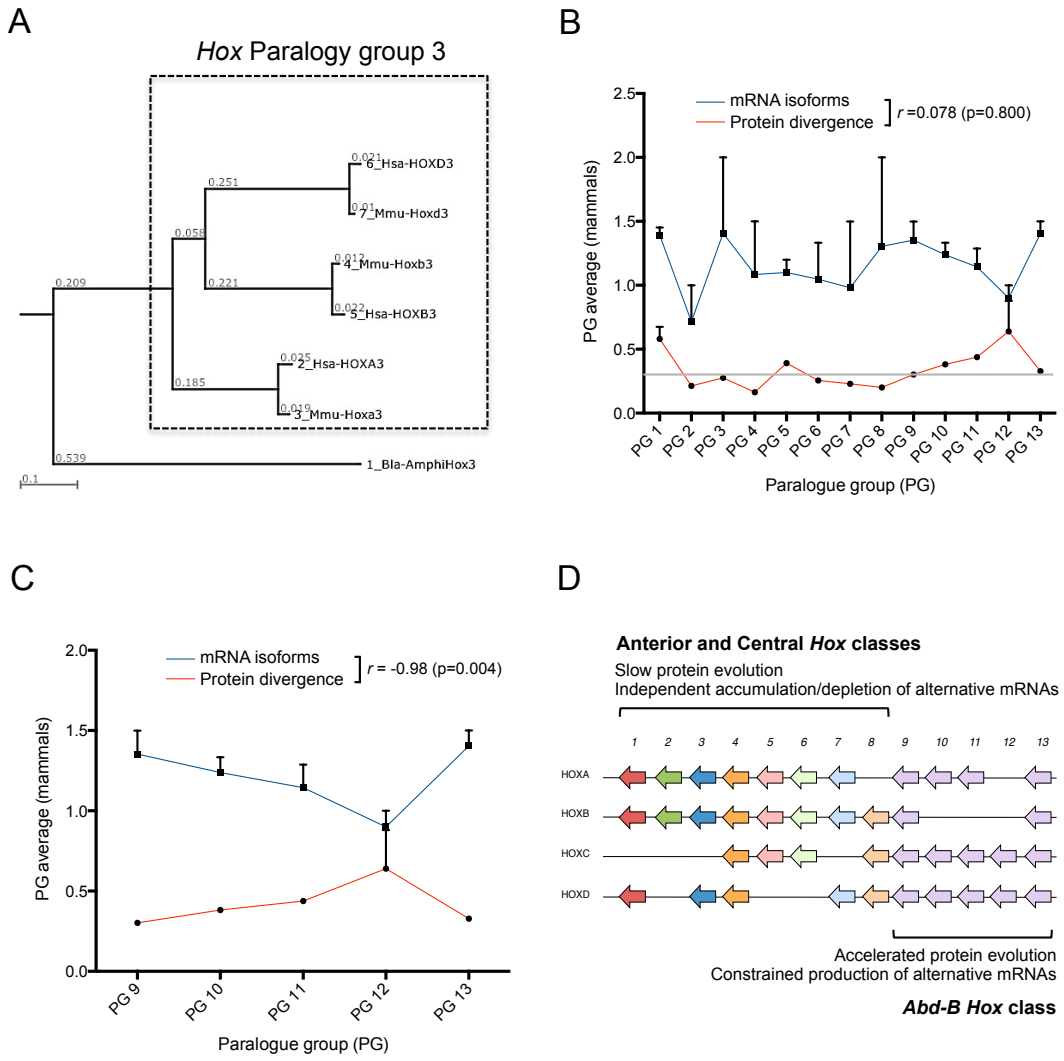


Figure 3.2 – Accelerated protein divergence within posterior *Hox* PGs is strongly associated with the reduced production of alternative mRNAs (legend in the following page)**.**

Figure 3.2 – Accelerated protein divergence within posterior *Hox* PGs is strongly associated with the reduced production of alternative mRNAs. (A) Example of the phylogenetic analysis used to calculate protein divergence rates within paralogue groups. The longest protein-coding isoforms from Hox3 paralogues were aligned with an out-group, the *Amphioxus* Hox3 protein, using the MAFFT alignment software (Katoh & Standley 2013). This alignment was then used to generate a phylogenetic tree using the Neighbour-Joining method. The same tree was used to calculate within-PG divergence rates in each mammalian species, by adding the tree distances between all paralogues after the split with the out-group, and dividing the number of amino acid substitutions per site by the total amount of paralogues in each group (3 in this example). This yielded an average amount of protein divergence for each PG in *Mus musculus* and *Homo sapiens*. **(B)** Comparison between protein divergence rates and average mRNA production in mammalian *Hox* paralogue groups. I see that protein divergence rates are uncorrelated with the average production of alternative mRNAs across mammalian *Hox* PGs (r=0.078, p=0.800). (**C)** Protein divergence rates are negatively correlated with average alternative isoform production in the posterior *Hox* paralogue groups of mammals. I find that in PGs *Hox9-13* there is a significant negative correlation between protein divergence rates after gene duplication and the amount of alternative mRNAs produced (r=-0.98, p=0.004). **(D)** Diagram summarising the results in panels (B-C). In PGs *Hox1-8*, slow rates of post-duplication protein evolution are uncorrelated with the production of alternative mRNAs. In PGs *Hox9-13* however, the rates of protein diverge and differential mRNA production are not independent, indicating that high levels of protein evolution could impose a constraint on the accumulation of differential mRNAs of posterior *Hox* genes, or vice-versa.

within a PG, while at the same time justifying the maintenance of highly related genes in the genome. Thus, I propose an *isoform-sharing* model, by which closely related genes work in conjunction to produce a diverse expression output increasing, in turn, the evolutionary pressure for the retention of individual paralogues as predicted by the DDC model.

In summary, I observe, first, that the production of alternative RNAs by mammalian *Hox* loci is comparable to other *loci* only when in the context of gene duplication; second, that some paralogue groups (notably *Hox3*) seem to have a higher incidence of DRP than others; third, I see that the differential pattern of *Hox* DRP across paralogue groups is significantly conserved across vertebrates; fourth, I observe that for posterior *Hox* genes, the average amount of protein-coding transcripts produced per *Hox* gene is negatively correlated with the evolutionary divergence of paralogous *loci*, indicating that there is an interplay between the amount of differential RNA production of *Hox* genes (conserved across vertebrates) and the rates of *Hox* protein evolution after gene duplication.

3.2.2 – A catalogue of mammalian alternative *Hox* RNA processing.

In the previous section, I describe the rates of differential RNA processing in the mammalian *Hox* clusters, and propose that these are closely linked to the evolution of *Hox* genes. What are, however, the *kinds* of differential RNA processing that *Hox* genes undergo? What are the RNA processing mechanisms that are involved in the production of alternative *Hox* mRNAs? In a transcriptome-wide study, Wang and colleagues (Wang et al. 2008) found that alternative splicing and alternative polyadenylation are

correlated in human tissues. Do I see that different kinds of DRP events also work in coordination to produce alternative *Hox* isoforms? To answer these questions, I needed to obtain a qualitative view of alternative mRNA processing in mammals. To achieve this, I first retrieved the nucleotide sequences for all *Hox* mRNAs (see the previous section). I then defined the reference isoform for each of the 39 *loci* in both mice and humans as the longest mRNA that encodes for a homeodomain, and aligned all alternative isoforms from a given locus to its reference mRNA. Finally, I determined the types of alternative transcript events that occur in mammalian *Hox* genes using the categorization of DRP events adopted by *Wang et al*. in their study of alternative isoform production and regulation in human tissue transcriptomes (Wang et al. 2008). There, eight distinct types of DRP events are defined: Skipped Exons (SE), Retained Introns (RI), Alternative 5' Splice-Sites (A5SS), Alternative 3' Splice-Sites (A3SS), Mutually exclusive exons (MXE), Alternative first exons (AFE), Alternative Last exons (ALE) and Tandem 3'UTRs (T3UTR) (see right panel in **Figure 3.4B**). I added a ninth category to these events, Tandem transcription start-sites (tTSS), as many *Hox* mRNA isoforms differed from the reference mRNA in the exact start of their transcriptional unit, a phenomenon that affects mRNA sequences but is not included in the AFE transcript event category, or indeed any other, in the Wang et al. (Wang et al. 2008) categorization.

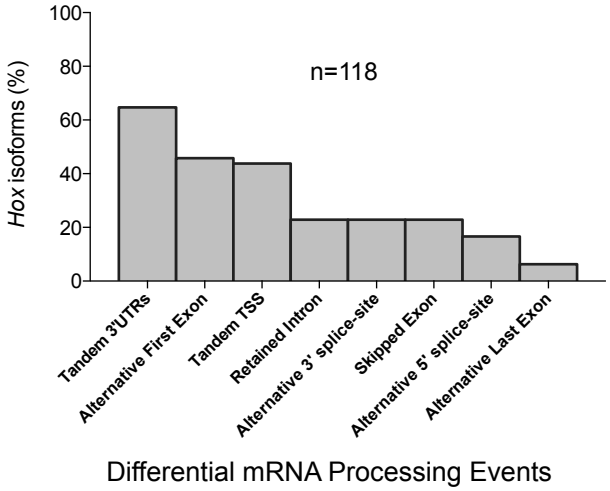
I consider the aforementioned parameters for the choice of a reference isoform (the longest protein coding and homeodomain-containing mRNA isoform of a *Hox* locus, see Chapter 2) to be reasonable. Furthermore, our reference isoforms coincide with the principal isoforms annotations of the APPRIS database in most cases. However, the choice of *reference vs. alternative* is necessarily arbitrary without evolutionary or tissue expression considerations; this is of concern, as the differential

RNA production of the *reference* isoform is not itself probed, an aspect of our analysis that could in principle lead to the loss of relevant information about *Hox* DRP. To control for this, I chose to annotate the occurrence of differential RNA processing events between the reference and the alternative isoforms with no regard for polarity, *e.g.* for a given *Hox* locus, an intron retention event that occurs in the alternative isoform but not in the respective reference mRNA is tabulated in the same manner (given a value of "1") if the intron were retained in the reference isoform and excised in the alternative *Hox* mRNA.

I find that the 47 alternative *Hox* mRNAs show a total of 118 individual DRP events when compared to the 31 reference *Hox* mRNAs (**Figure 3.3A**). The most flagrant DRP event is the generation of alternative 3'UTRs in tandem (t3UTR) in 66% of transcripts. This event is followed in preponderance by tTSS and AFE events (44.7% and 46.8%, respectively); the remaining categories occur at a rate of 25% or less. No MXE event was observed, and this DRP category was therefore removed from further analyses. These results suggest that the concerted remodelling of both 5'UTRs and 3'UTRs, through the control of alternative transcription start and termination sites, is the most common outcome of *Hox* DRP in mammals.

I next wondered whether these RNA processing events had an effect on the open-reading frames (ORF) sequences of each alternative *Hox* mRNA, and could lead to alternative *Hox* proteins. In order to study this aspect, I compared all alternative ORFs to the ORF present in the reference mRNA of each locus; I find that *Hox* DRP leads to alternative open-reading frames in 64% of the cases, indicating that RNA processing significantly expands *Hox* proteomic diversity. Together these results indicate that i) multiple DRP events intervene in the production of most alternative *Hox* transcripts, leading to the possibility that some events co-occur, ii) that alternative

A



B

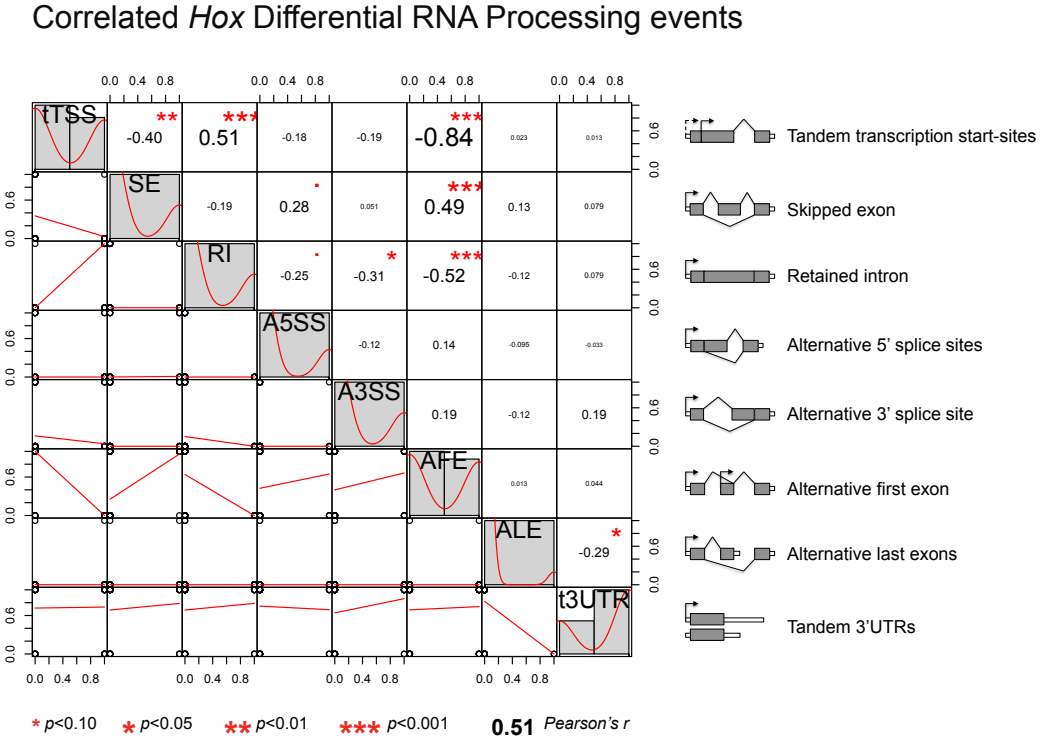


Figure 3.3 – Specific modes of differential *Hox* RNA processing show distinct links to transcriptional regulation (legend in the following page)**.**

Figure 3.3 – Specific modes of differential *Hox* RNA processing show distinct links to transcriptional regulation. (A) Mammalian *Hox* genes undergo different kinds of differential RNA processing event. Using the tabulation of differential RNA processing events in (Wang et al. 2008), and adding the occurrence of tandem Transcription Start Sites (tTSSs), I find a total of 118 differential RNA processing events involved in the formation of alternative *Hox* mRNAs. The most represented differential RNA processing event is the formation of alternative 3'UTRs in tandem, occurring in 66% of the cases, followed by the occurrence of Alternative First Exons (AFEs) and tTSSs in 46.8% and 44.7% of the cases, respectively. I find no occurrence of Mutually-Exclusive Exons (MXEs). This category was excluded from further analyses. **(B)** Specific modes of differential Hox RNA processing show distinct links to transcriptional regulation. I find that the usage of tTSSs upon transcriptional initiation is positively correlated with the occurrence of alternative splicing by the Retention of Introns (RI, r= 0.51, p=0.0002), and negatively correlated with the occurrence of Skipped Exons (SE, r= -0.40, p=0.0059). Conversely, I see that AFEs are negatively correlated with RI (r= -0.52, p=0.0002) and positively correlated with SE (r=0.49, p=0.0005). The most represented DRP event, tandem 3'UTR formation, is uncorrelated with other RNA processing events. These results suggest that the differential RNA processing of mammalian *Hox* genes includes at least two distinct links between the regulation of transcription and alternative splicing.

transcription and alternative transcriptional termination through alternative 3'UTR cleavage and polyadenylation (collectively deemed APA from here on) are the most prevalent processing events that intervene in differential *Hox* RNA processing and iii) that protein-coding sequences are more commonly than not affected by *Hox* DRP.

Given that most *Hox* mRNAs show evidence of more than one type of DRP event, I next considered the possibility that some DRP events might co-occur more often than others, so as to find evidence supporting the hypothesis that the production of *Hox* mRNAs requires the coordination of different RNA processing mechanisms. In order to answer this question, I tabulated the DRP event data for all 47 alternative *Hox* isoforms, granting the numerical values "1" and "0", respectively, to the presence or absence of each DRP event on each alternative *Hox* transcript. I next submitted this binary table into the statistical analysis software *R*, and used the *PerformanceAnalytics R* package to ask whether DRP events were significantly correlated to each other in the generation of alternative *Hox* transcripts.

I find that the two mutually exclusive categories of transcriptional initiation analysed, tandem Transcription start sites (tTTS), which sit close together in the genome and are thus controlled in all probability by the same transcriptional promoter (see Chapter 1), and alternative first exons (AFE), which denote transcription initiation sites that are far apart and could thus be controlled by different promoters, both correlate with alternative splicing categories (**Figure 3.3B**). However, while tTTSs correlate negatively with exon skipping (r=-0.40, p=0.0059), AFEs correlate positively with the same alternative splicing DRP (r=0.49, p=0.0005) (**Figure 3.3B**). Conversely, intron retention (RI) AS events are positively correlated with tTTSs (r=0.51, p=0.0002), and negatively correlated with AFE events (r=-0.52, p=0.0002). This indicates that the alternative control of *Hox* transcriptional initiation can impact the alternative splicing

patterns of the nascent mRNA, a phenomenon that has been previously observed in a number of cases (Kornblihtt et al. 2004) (**Figure 3.3B)**.

In terms of the coordination between different alternative splicing events, I see a significant negative correlation of medium effect between Intron Retention and alternative 3'splice-site usage (r = -0.31, p=0.0367) (**Figure 3.3B)**. This indicates that once an intron is retained alternative 3'splice-sites are less prone to be used in downstream DRP events. I hypothesize that the mechanistic explanation for the link between these two patterns could lie in the competition between splice-sites for a *trans*-regulator. Finally, I find no correlation between the most represented DRP event, tandem 3'UTR production, and other DRP event. The other DRP event that leads to alternative 3'UTRs in our dataset, ALE, also appears uncorrelated with other DRP events in the transcripts analysed (**Figure 3.3B)**. This suggests that 3'end formation can happen somewhat independently of other upstream DRP operations. These results suggest that *Hox* differential RNA processing involves the coordination between specific types of transcription and alternative splicing, and that 3'UTRs are often remodelled independently of previous differential RNA processing operations.

To obtain a picture of these multiple coordinated DRP events at the transcript level, and identity which genes undergo different types of coordinated DRP, I hierarchically clustered the previously mentioned binary matrix of *Hox* DRP events per transcript using the R function *hclust* (see Chapter 2). I find that coordinated processing occurs in all but 6 *Hox* transcripts (**Figure 3.4A).** Moreover, the hierarchical clustering of similarly processed alternative *Hox* RNAs reveals two main statistically supported clusters of transcripts with similar DRP profiles, which I will hereafter call α and β (**Figure 3.4)**. In cluster α, all 22 transcripts are generated by alternative transcription leading to alternative first exons, and then undergo many different types of

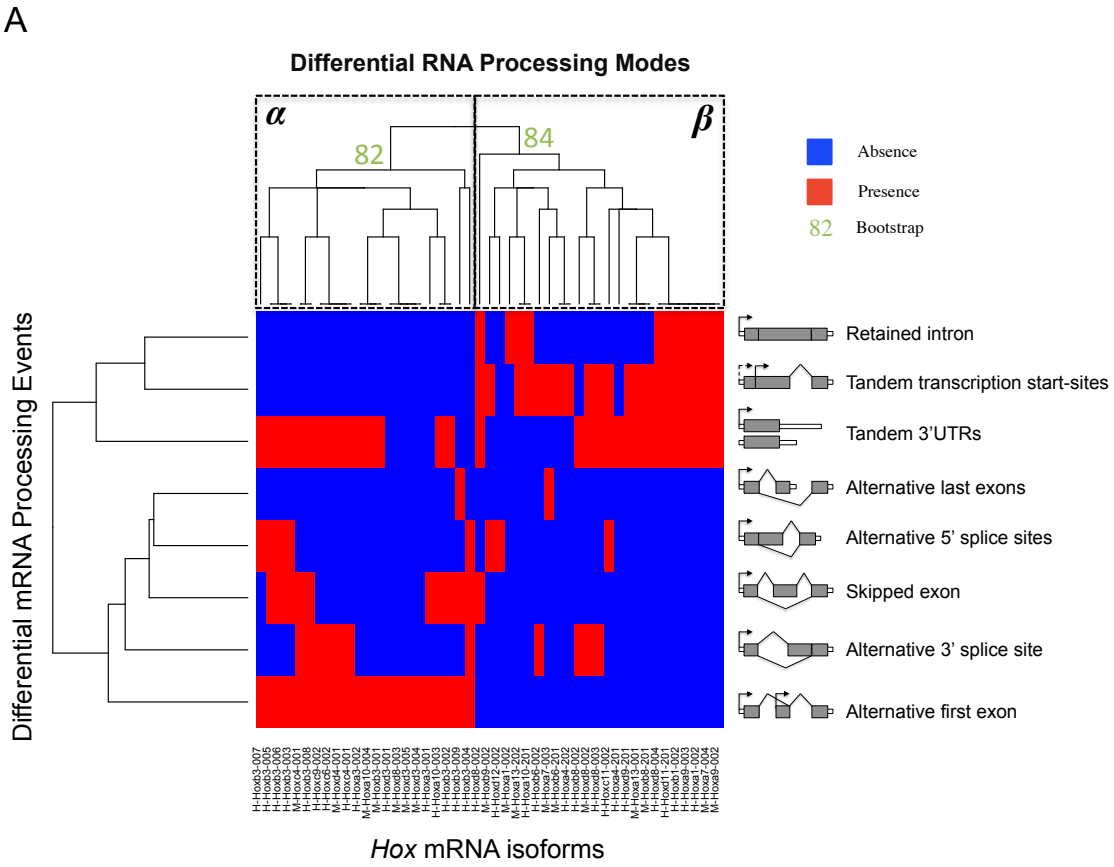


Figure 3.4 – Differential *Hox* RNA processing involves the coordination of multiple regulatory levels and two distinct modes (legend in the following page)**.**

Figure 3.4 – Differential *Hox* RNA processing involves the coordination of multiple regulatory levels and two distinct modes. (A) Hierarchical clustering of differential RNA processing events involved in the production of mammalian *Hox* mRNAs. Hierarchical clustering of all alternative *Hox* mRNAs based on the occurrence or absence of each individual DRP event (right panel). I observe two well-supported clusters of alternative *Hox* mRNAs, α and β, which involve the coordination of different kinds of DRP events. In cluster α, all 22 transcripts are generated by alternative transcription leading to alternative first exons, and then undergo many different types of alternative splicing, mostly exon skipping. In category β, most *Hox* transcripts show evidence of tandem transcription start-sites (tTSS) and alternative splicing by intron retention. Tandem 3'UTRs, the most represented DRP event in our dataset, features heavily in both clusters, indicating that this is a general theme in the production of alternative *Hox* mRNAs. Interestingly, all mammalian isoforms of PG 3 fall into cluster α, as well as most isoforms from PGs 4 and 10. Conversely, all isoforms of paralogue groups 1, 7, 11, 12 and 13 fall into category β, as well as most isoforms generated by PGs 8 and 9. This indicates that in at least 10 of the 13 PGs, DRP modes are shared across PGs. **(B)** Diagram summarising the finding in panel A. The differential RNA processing of mammalian *Hox* mRNAs proceeds by at least two distinct modes, α and β, which coordinate transcriptional regulation, alternative splicing, and alternative cleavage and polyadenylation in a paralogous-specific manner.

alternative splicing, mostly exon skipping; only in 10 out of 22 cases does DRP remodel protein-coding sequences. Additionally, most transcripts undergo alternative 3'UTR formation in tandem. In category β, most of the 25 alternative *Hox* transcripts are generated by tandem transcription start-sites close by (tTSS), and then undergo alternative splicing, mostly through intron retention. Protein-coding sequences are affected in 20 of the 25 cases. As with category α, 3'UTRs are mostly remodelled in tandem. Interestingly, all isoforms of paralogue group 3, in both mice and humans, fall into category α, as well as most isoforms from PGs 4 and 10. Conversely, all isoforms of paralogue groups 1, 7, 11, 12 and 13 fall into category β, as well as most isoforms generated by PGs 8 and 9. These results indicate that the general patterns of coordinated *Hox* RNA processing are mostly shared by paralogous genes, indicating two non-exclusive possibilities: the *cis*-regulatory sequences that mediate these processes were already present before gene duplication, and/or the shared expression patterns of many paralogous *Hox* genes leads to the exposure of paralogous *Hox* mRNAs to the same molecular environments and thus to the same *trans*-regulators of DRP, leading to the similar DRP patterns.

3.2.3 – The alternative 3'UTRs of mammalian *Hox* mRNAs show a conserved segregation of microRNA (miRNA) target-sites.

In the previous section, I report that the formation of alternative 3'UTRs in tandem is the single most represented *Hox* DRP event, occurring in 66% of alternative mRNA production cases. Indeed, of the 18 *Hox* genes that show alternative isoforms in humans, 15 form alternative 3'UTRs (9 out of 13 in the case of mice). This result points

to the prospect that the ability to produce mRNA with different 3'UTR sequences is a staple of the regulation of *Hox* gene expression in mammals. One clear hypothesis that results from this data is that alternative *Hox* transcripts regulate their *visibility* to RNA regulators in *trans* through the inclusion or exclusion of *cis*-regulatory sequences in the 3'UTRs (Thomsen et al. 2010). One corollary of this hypothesis is that *cis*-regulatory complements are different when alternative 3'UTRs of the same gene are compared.

I thus wanted to test whether the different 3'UTRs of the same genes carried similar or different sequences, and possibly regulatory information. Regulatory sequences in *cis* have been previously shown to mainly mediate the interaction of the 3'UTRs of mRNAs with both miRNAs and RNA-Binding Proteins (RBPs). The latter are proteins with RNA-binding domains, known to bind mRNAs directly and mediate the regulation of mRNA processing itself, as well as mRNA nuclear export, localization, stability and translation rates through their interactions with mRNAs as well as other proteins and noncoding RNAs. However, few studies implicate RBPs in the differential regulation of *Hox* alternative isoforms, one notable exception being the study of alternative *Ubx* RNA processing by the RBP ELAV in the CNS of *Drosophila melanogaster* (See Chapters 1 and the final section of the current Chapter). In contrast, a significant number of studies indicate that miRNA-based regulation is very prevalent in *Hox* genes. I chose miRNAs as candidate *trans*-regulators of *Hox* mRNAs for a number of reasons. These molecules are computationally predicted to target a third of human genes, and have been shown to target mammalian *Hox* genes in many cases (see miRTarBase (Hsu et al. 2014) and **Table 3.1**).

Given the previous reports of miRNA-mediated regulation of *Hox* gene expression (see Chapter 1), and the importance of alternative 3'UTR sequence contexts to the outcome of this regulatory mechanism, I wondered whether there was a

segregation of miRNA target complements between different portions of mammalian *Hox* 3'UTRs. To test this hypothesis, I retrieved a list of miRNA loci experimentally shown to mediate the down-regulation of *Hox* mRNAs, and studied where their target sites were in the context of the observed alternative 3'UTR formation in *Hox* genes.

I used miRNA targeting prediction tool PITA (Kertesz et al. 2007) to predict miRNA targets for the all the *Hox* genes with alternative 3'UTR formation, performing separate, species-specific miRNA targeting predictions for the constitutive and optional 3'UTRs (hereafter deemed "Short" and "Long" 3'UTRs, respectively). PITA uses a thermodynamic approach to predict miRNA targets in the 3'UTRs of mRNAs. It first predicts the energy gained by a miRNA-mRNA pair upon complementary binding, and then subtracts from this value the energy needed to undo local target mRNA structure as a consequence of internal mRNA base pairing which RNAFold predicts. This computation results in a measurement, deemed "ΔΔG", that decreases in numerical value as the predicted miRNA targeting efficiency increases.

I then filtered miRNA-targeting predictions so that these included only experimentally validated miRNA-targeting events in the *Hox* genes of either *Homo sapiens* or *Mus musculus*. The miRTarBase online repository (http://mirtarbase.mbc.nctu.edu.tw) documents 86 experimentally-validated interactions between specific mature miRNAs and the 15 *Hox* genes that show alternative 3'UTRs in humans, as well as the experimental method used to assess this molecular interaction (see **Table 3.1**). In the case of *Mus musculus*, only 4 miRNA-*Hox* interactions were reported (see **Table 3.1**). Given the reduced number of experimentally validated miRNA-*Hox* interactions for *Mus musculus*, I decided to use the *M. musculus* homologues of human *Hox*-targeting miRNAs in subsequent analyses. I additionally filtered results to include only miRNAs that are present in mice and humans, as well as

Table 3.1 – Experimental techniques used in the validation of mammalian Hox-miRNA interactions (data retrieved from miRTarBase).

ID	Species (miRNA)	Species (Target)	miRNA	Target	Validation methods							Sum	Papers
					Strong evidence			Less strong evidence					
					Reporter Assay	Western Blot	qPCR	Microarray	NGS	pSILAC	Other		
MIRT006838	Homo sapiens	Homo sapiens	hsa-miR-7-5p	**HOXB3**	✓							1	1
MIRT006839	Homo sapiens	Homo sapiens	hsa-miR-218-5p	**HOXB3**	✓							1	1
MIRT025627	Homo sapiens	Homo sapiens	hsa-miR-10a-5p	**HOXB3**					✓			1	1
MIRT000190	Homo sapiens	Homo sapiens	hsa-miR-204-5p	**HOXA10**		✓		✓				2	2
MIRT004850	Homo sapiens	Homo sapiens	hsa-miR-192-5p	**HOXA10**	✓		✓					2	1
MIRT005441	Homo sapiens	Homo sapiens	hsa-miR-130a-3p	**HOXA10**	✓		✓	✓				3	1
MIRT006793	Homo sapiens	Homo sapiens	hsa-miR-135a-5p	**HOXA10**	✓							1	1
MIRT024822	Homo sapiens	Homo sapiens	hsa-miR-215-5p	**HOXA10**				✓				1	1
MIRT044613	Homo sapiens	Homo sapiens	hsa-miR-320a	**HOXA10**					✓			1	1
MIRT045606	Homo sapiens	Homo sapiens	hsa-miR-149-5p	**HOXA10**					✓			1	1
MIRT002941	Homo sapiens	Homo sapiens	hsa-miR-196a-5p	**HOXD8**	✓	✓					✓	3	3
MIRT000149	Homo sapiens	Homo sapiens	hsa-miR-210-3p	**HOXA9**	✓						✓	2	1
MIRT001016	Mus musculus	Mus musculus	mmu-let-7a-5p	**Hoxa9**		✓					✓	2	1
MIRT001017	Mus musculus	Mus musculus	mmu-miR-145a-5p	**Hoxa9**		✓					✓	2	1
MIRT001917	Homo sapiens	Homo sapiens	hsa-miR-145-5p	**HOXA9**	✓	✓	✓				✓	4	1
MIRT001918	Homo sapiens	Homo sapiens	hsa-miR-126-3p	**HOXA9**	✓	✓	✓				✓	4	1
MIRT002288	Mus musculus	Mus musculus	mmu-miR-126a-3p	**Hoxa9**		✓					✓	2	1
MIRT016142	Homo sapiens	Homo sapiens	hsa-miR-652-3p	**HOXA9**				✓				1	1
MIRT019591	Homo sapiens	Homo sapiens	hsa-miR-340-5p	**HOXA9**				✓				1	1
MIRT027337	Homo sapiens	Homo sapiens	hsa-miR-101-3p	**HOXA9**				✓				1	1
MIRT027927	Homo sapiens	Homo sapiens	hsa-miR-96-5p	**HOXA9**				✓				1	1
MIRT028225	Homo sapiens	Homo sapiens	hsa-miR-33a-5p	**HOXA9**				✓				1	1
MIRT030092	Homo sapiens	Homo sapiens	hsa-miR-26b-5p	**HOXA9**				✓				1	1
MIRT030986	Homo sapiens	Homo sapiens	hsa-miR-21-5p	**HOXA9**			✓					1	1
MIRT031333	Homo sapiens	Homo sapiens	hsa-miR-18a-5p	**HOXA9**				✓				1	1
MIRT039078	Homo sapiens	Homo sapiens	hsa-miR-769-3p	**HOXA9**				✓				1	1
MIRT043225	Homo sapiens	Homo sapiens	hsa-miR-324-5p	**HOXA9**				✓				1	1
MIRT000152	Homo sapiens	Homo sapiens	hsa-miR-210-3p	**HOXA1**	✓						✓	2	1
MIRT001140	Homo sapiens	Homo sapiens	hsa-miR-10a-5p	**HOXA1**	✓	✓	✓				✓	4	2
MIRT017144	Homo sapiens	Homo sapiens	hsa-miR-335-5p	**HOXC6**				✓				1	1
MIRT021916	Homo sapiens	Homo sapiens	hsa-miR-128-3p	**HOXC6**					✓			1	1
MIRT028721	Homo sapiens	Homo sapiens	hsa-miR-27a-3p	**HOXC6**					✓			1	1

adapted from data in **miRTarBase** (http://mirtarbase.mbc.nctu.edu.tw)

removing positive ΔΔG values (miRNA-target interactions predicted to be energetically implausible).

I find that in humans, most experimentally validated miRNA-*Hox* interactions have predicted miRNA targets that lie in the distal, rather than the proximal tract of *Hox* 3'UTRs (20 and 9 miRNA targets, respectively) (**Figure 3.5A**). In the case of mice, this is also true (11 targets in distal 3'UTRs, 8 in proximal 3'UTRs) - (**Figure 3.5B**). These results could be explained by the fact that the distal tracts of *Hox* 3'UTRs are on average longer than the respective constitutive regions. However, I also see that miRNA targets for validated miRNA-*Hox* interactions have a lower ΔΔG value, and are thus stronger on average when in the context of the distal 3'UTR (**Figures 3.5C-D**). In the case of *Homo sapiens*, the difference in targeting between proximal and distal 3'UTRs is statistically significant (**Figure 3.5C**), and while this is not so for *Mus musculus* (**Figure 3.5D**), the average targeting is even larger for the distal 3'UTR, when compared to humans, leading us to think that the relative lack of miRNA targeting data in *Mus musculus* could be responsible for the aforementioned lack of statistical significance.

These results point to a bias in the positioning of biologically relevant miRNA targets within *Hox* 3'UTRs. However, this dataset is hardly exhaustive and is biased, as it depends on remotely validated miRNA-mRNA interactions performed by various independent research programs. I expect that it is due to this that I observe the lack of miRNA-mRNA interaction data for most *Hox* genes in mice, as well as some in humans, where alternative 3'UTR formation is present. Friedman and colleagues (Friedman et al. 2009) have used an evolutionary approach based on target-site conservation to show that mammalian 3'UTR carry on average 4.2 sites for miRNAs. This stands in contrast to our analysis, where I find 2.5 experimentally validated

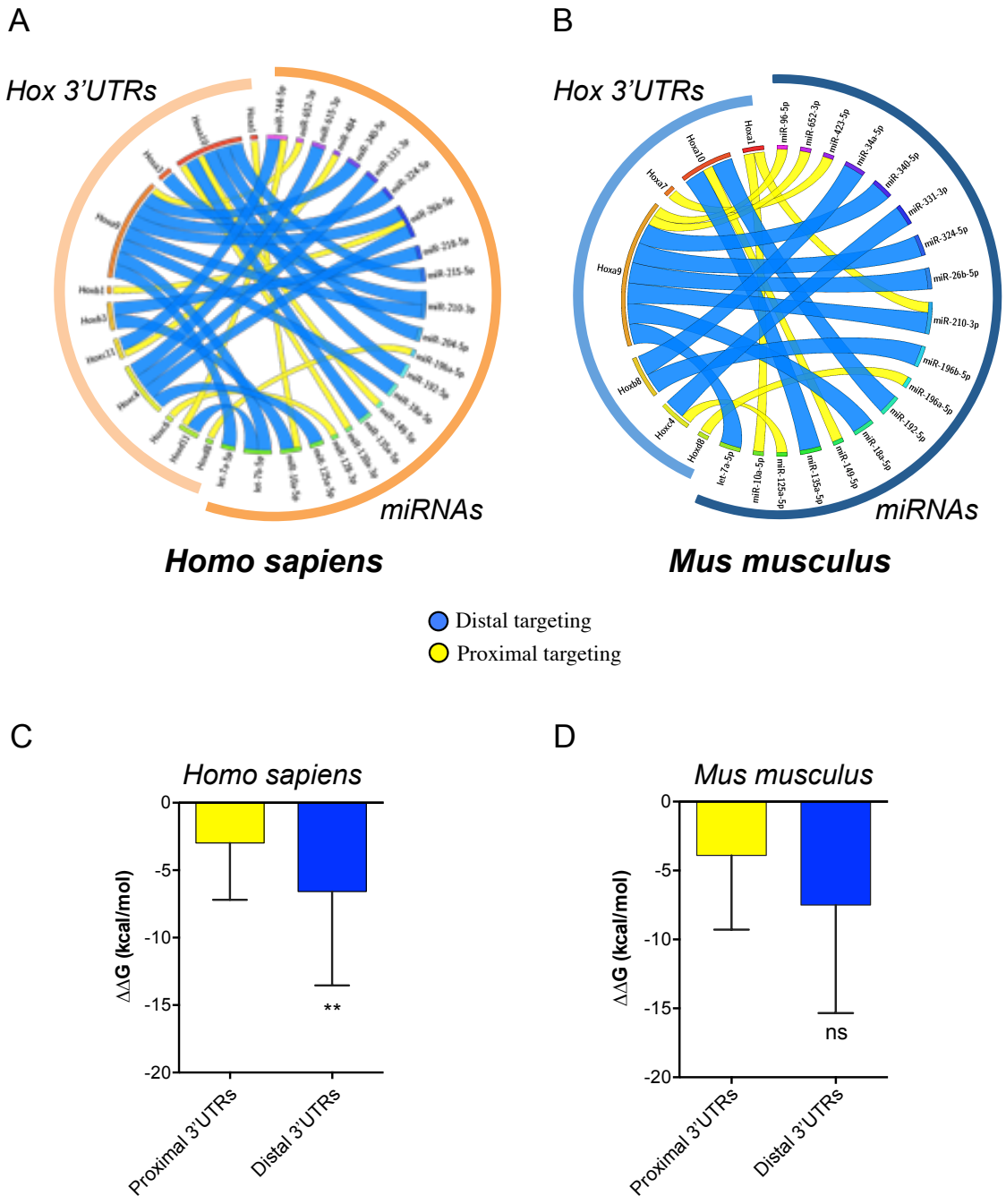


Figure 3.5 – Experimentally validated miRNAs are predicted to bind to more numerous and stronger targets in distal *Hox* 3'UTRs (legend in the following page)**.**

Figure 3.5 – Experimentally validated miRNAs are predicted to bind to more numerous and stronger targets in distal *Hox* 3'UTRs. (A-B) Most experimentally validated miRNA-*Hox* interactions are predicted to occur in the distal (blue), rather than the proximal (yellow) 3'UTRs of *Hox* genes in both (A) *Homo sapiens* and (B) *Mus musculus*. **(C-D)** Distal 3'UTRs of *Hox* genes (blue) contain stronger miRNA target sites than their proximal counterparts (yellow). All experimentally-validated miRNA-*Hox* interactions were retrieved from miRTarBase (Hsu et al. 2014); miRNA targeting predictions were performed using PITA (Kertesz et al. 2007). The experimental techniques used in the validation of miRNA-*Hox* targeting can be found in **Table 3.1**. The results of the miRNA targeting predictions used in these analyses can be found in **Table 3.2**.

Table 3.2 - 3'UTR targeting predictions of experimentally validated miRNA-*Hox* interactions using PITA.

Homo sapiens		ΔΔG (*kcal mol* $^{-1}$)	
Gene	*microRNA*	Distal 3'UTRs	Short 3'UTRs
Hoxa9	*hsa-miR-324-5p*	-26,91	0
Hoxc11	*hsa-miR-744-5p*	-20,8	-13,19
Hoxa9	*hsa-miR-210-3p*	-15,71	0
Hoxb3	*hsa-miR-10a-5p*	-13,7	0
Hoxa9	*hsa-miR-18a-5p*	-12,81	-6,6
Hoxa9	*hsa-let-7a-5p*	-11,82	0
Hoxa9	*hsa-miR-26b-5p*	-11,2	0
Hoxa9	*hsa-miR-340-5p*	-9,74	0
Hoxd11	*hsa-let-7b-5p*	-8,3	0
Hoxa10	*hsa-miR-192-5p*	-8,22	0
Hoxa10	*hsa-miR-135a-5p*	-7,82	0
Hoxd11	*hsa-miR-744-5p*	-7,52	-14,04
Hoxc4	*hsa-miR-615-3p*	-7,36	0
Hoxa10	*hsa-miR-215-5p*	-7,32	-1,89
Hoxb3	*hsa-miR-218-5p*	-6,4	-4,9
Hoxc4	*hsa-miR-125a-5p*	-4,71	0
Hoxc4	*hsa-miR-331-3p*	-1,68	0
Hoxc4	*hsa-miR-26b-5p*	-1,3	0
Hoxa3	*hsa-miR-210-3p*	-0,78	0
Hoxa10	*hsa-miR-204-5p*	-0,11	0
Hoxa10	*hsa-miR-130a-3p*	0	-6,91
Hoxa10	*hsa-miR-149-5p*	0	-5,18
Hoxd8	*hsa-miR-196a-5p*	0	-4,62
Hoxa1	*hsa-miR-10a-5p*	0	-3,78
Hoxc6	*hsa-miR-128-3p*	0	-1,54
Hoxc11	*hsa-miR-484*	0	-2,66
Hoxa9	*hsa-miR-652-3p*	0	-11,35
Hoxb1	*hsa-miR-26b-5p*	0	-6,74

Mus musculus		ΔΔG (*kcal mol -1*)	
Gene	*microRNA*	Distal 3'UTRs	Proximal 3'UTRs
Hoxb8	*mmu-miR-196b-5p*	-25,59	-0,15
Hoxa9	*mmu-miR-324-5p*	-20,1	0
Hoxa9	*mmu-miR-210-3p*	-16,04	-9,83
Hoxb8	*mmu-miR-34a-5p*	-12,66	0
Hoxa9	*mmu-miR-18a-5p*	-12,33	-4,98
Hoxa9	*mmu-miR-652-3p*	-12,3	-15,86
Hoxa9	*mmu-miR-26b-5p*	-12,1	0
Hoxa9	*let-mmu-7a-5p*	-11,14	0
Hoxa9	*mmu-miR-340-5p*	-9,04	0
Hoxa10	*mmu-miR-135a-5p*	-5,85	0
Hoxa10	*mmu-miR-192-5p*	-2,7	0
Hoxc4	*mmu-miR-331-3p*	-1,65	0
Hoxa10	*mmu-miR-149-5p*	-0,88	-6,58
Hoxa1	*mmu-miR-10a-5p*	0	-14,7
Hoxd8	*mmu-miR-196a-5p*	0	-9,97
Hoxc4	*mmu-miR-125a-5p*	0	-8,51
Hoxa7	*mmu-miR-423-5p*	0	-0,74
Hoxa1	*mmu-miR-210-3p*	0	-0,1
Hoxa9	*mmu-miR-96-5p*	0	-2,56

miRNA targets, on average, in human *Hox* 3'UTRs - our most enriched dataset – while at least half of the *Hox* genes analysed have a single miRNA-mRNA interaction. This comparison indicates that our analyses include a conservative number of miRNA-mRNA interactions, suggesting that an expansion of our dataset might be of use in further analyses.

As these results suggest that there is indeed a conserved segregation of miRNA target-sites between short and long 3'UTR portions of *Hox* mRNAs, but are expected to be an incomplete picture of miRNA-mediated *Hox* regulation, I next decided to analyse the *Hox* 3'UTR targeting predictions for all conserved mammalian miRNAs, with the goal of seeing if, first, the observed tendency could be extended to the entire miRNA targeting complement for each gene, and second, if this tendency was statistically significant. To this end, I first identified all miRNAs that are conserved in both mammalian species using the miRNAminer tool (Artzi et al. 2008) (http://groups.csail.mit.edu/pag/mirnaminer/), and then retrieved the mature sequences of all miRNAs produced by these loci from miRBASE (Kozomara & Griffiths-Jones 2013) (http://www.mirbase.org). I decided not to use miRNAs that have duplicated in either of the two mammalian lineages analysed, as the exact homology status of each paralogous miRNA was uncertain. Using the aforementioned approach, I isolated a set of 438 mature miRNAs that are unambiguously conserved between mice and humans. As the specific mature miRNA sequences show, in some cases, a small degree of divergence across mammals, I decided to use the species-specific variants for each miRNA in all miRNA predictions. This approach is expected to control for miRNA-target co-evolution (Barbash et al. 2014), and thus presents a good opportunity to study the evolution of miRNA targeting in relation to *Hox* tandem 3'UTR formation.

With the expectation that a broader dataset for *Hox*-miRNA interactions might

support and expand our initial observation, that APA introduces differential miRNA target complements across alternative mRNA isoforms of the same *Hox* gene, I used PITA and each species' variant of the mammalian miRNA complement to predict all miRNA target sites in both constitutive and elective 3'UTRs. As most of these predicted interactions are not experimentally validated, I decided to introduce one stringency criterion in order to decrease the incidence of false-positives in our analysis: I used the authors' recommended threshold of $\Delta\Delta G < -10$ (Kertesz et al. 2007), so as to have a more conservative dataset that included only the strong miRNA-target complement of each portion of *Hox* 3'UTRs (see Chapter 2). In the case of *Hoxb8*, most 3'UTRs did not present miRNA targets with such low $\Delta\Delta G$ values; as such, I used miRNA $\Delta\Delta G < -8$ for the 3'UTRs of this gene. This is expected to add further confidence to the miRNA target predictions, as it allows us to compare not only the existence targeting but also its amount.

I next performed a miRNA target correlation analysis between Short and Long 3'UTRs of the same gene in each species. I analysed eight of the fifteen human *Hox* genes with tandem 3'UTRs, as the remaining two had 3'UTR tracts that were too short (less than 70 nucleotides in length) to be used by PITA. In these eight genes, I find that all display a significant correlation between Short and Long miRNA complements (**Figure 3.6A**). Interestingly, this correlation is negative in value, in most cases (six). This tendency is also observed in the *Hox* genes of mice (**Figure 3.6A**). These results further indicate that, within each species, the miRNA complement of the 3'UTRs is segregated across alternative isoforms for *Hox* genes with alternative 3'UTR formation in mammals.

As I used the human-mouse conserved miRNA complement in our predictions, I next compared orthologous 3'UTRs across species (e.g. *Hoxb3-Short* in

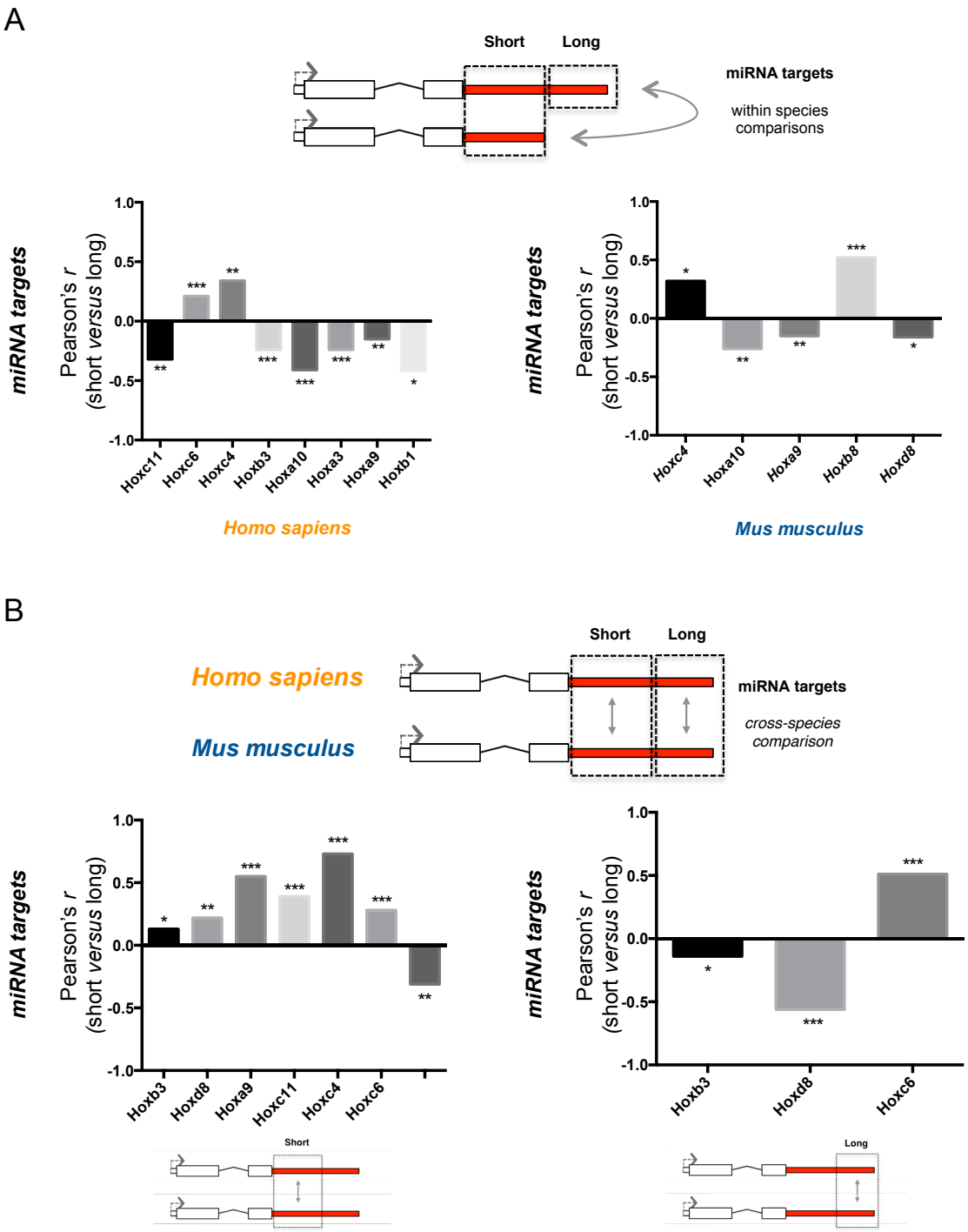


Figure 3.6 – Alternative cleavage and polyadenylation generates developmental and evolutionary compartments in *Hox* 3'UTRs (legend in the following page)**.**

Figure 3.6 – Alternative cleavage and polyadenylation generates developmental and evolutionary compartments in *Hox* 3'UTRs. (A-B) miRNA targeting predictions in the context of alternative 3'UTR formation in the mammalian *Hox* genes. (A) Comparison of miRNA targeting across alternative *Hox* 3'UTR isoforms of the same locus. I see that the miRNA target complement of Short and Long 3'UTRs of individual *Hox* loci is negatively correlated in the majority cases, in both *Homo sapiens* and *Mus musculus*. (B) Comparison of miRNA targeting across homologous *Hox* 3'UTR isoforms of different mammalian species. When I compare the miRNA targeting predictions for the Short 3'UTRs of orthologous *Hox* genes, I find that 3'UTR targeting is positively correlated across species (left panel), indicating that these sequences are conserved between *Homo sapiens* and *Mus musculus* with respect to miRNA regulation. Conversely, the comparison of miRNA targeting in the distal 3'UTRs yields negative correlations in two out of three cases, indicating that these *Hox* 3'UTR sequences have significantly diverged between humans and mice. All predictions were performed using PITA. Only miRNA targets with a $\Delta\Delta G \leq -10$ were used, as per the recommendation of the authors (Kertesz et al. 2007). I used the conserved mammalian miRNA complement in all analyses, as annotated in miRNAminer (Artzi et al. 2008) and miRBase (Kozomara & Griffiths-Jones 2013).

mice with *Hoxb3-Short* in humans). I find that unlike the within-species comparisons, the miRNA complement of a given gene's Short 3'UTRs is positively correlated across species in six out of seven cases (**Figure 3.6B**). A similar cross-species comparison of the "Long" 3'UTR tracts of the same gene yields negative correlations in two of three cases, *Hoxb3* and *Hoxd8* (r= -0.14* and r= -0.56***, respectively), (**Figure 3.6B**).

Together, these results indicate that the 3'UTR miRNA-target complements of *Hox* genes are not only segregated across 3'UTR isoforms for a given gene in a given species, but also that these distinct miRNA target complements evolve differently, being more conserved in Short than Long 3'UTRs. To explain these results, I propose that APA can generate both developmental and evolutionary compartments in mammalian *Hox* genes, with Short/constitutive 3'UTRs maintaining, on average, their miRNA targets across evolution, while the use of distal 3'UTR tracts introduces novel rather than redundant miRNA targets on an mRNA; essential and conserved miRNA targets would accumulate in the constitutive 3'UTR across evolution, while the miRNAs that are necessary for species-specific types of regulation in time and space would be segregated to the distal, optional, 3'UTR.

3.3 – Discussion.

During the course of this work, I have used RNA-sequencing data of both developing and adult *Strigamia maritima* centipedes, in the context of the genome and transcriptome sequencing project of this organism (Chipman et al. 2014). In that work, I show that posterior *S. maritima Hox* genes, homologous to *Hox* loci of the *Drosophila melanogaster* BX-C class, tend to produce differential 3'UTR isoforms (Chipman et al.

2014). This pattern is conserved between *Drosophila melanogaster* and *Strigamia maritima* (Thomsen et al. 2010; Chipman et al. 2014), suggesting that differential RNA processing is an important regulatory level in the regulation of *Hox* gene expression of arthropods.

In this chapter, I investigate the rates, quality and evolutionary dynamic of differential RNA processing in the mammalian *Hox* clusters, using the freely available GENCODE dataset. I start by comparing the absolute incidence of RNA processing across *Hox* clusters with genome-wide estimates in both the literature and the GENCODE database, and find that the average incidence of DRP on *Hox* genes is significantly smaller than expected in both *Homo sapiens* and *Mus musculus*, even when taking account the limitations of the database. I then show that this can be explained, in part by the duplication History of the *Hox* group, as I see that the average production of alternative mRNAs per gene within a paralogue group is consistent with the rest of the genome. This led us to hypothesize that there is functional sharing of isoforms between paralogous *Hox* genes.

The proposed functional sharing fits with the subfunctionalisation model of alternative splicing evolution after gene duplication (Su et al. 2006). According to this model, as the *Hox clusters* of early vertebrates underwent two rounds of duplication – concurrently with the rest of the genome –, the production of differential mRNA isoforms in the ancestral cluster is distributed among the resulting paralogues, with each duplicant retaining a portion of the ancestral *Hox* expression output (Kopelman et al. 2005; Su et al. 2006). As such, different paralogues fix, on average, different isoforms from the total pool of ancestral mRNAs; as such, the pool of alternative mRNAs from one ancestral *Hox* locus is maintained in the early stages after gene duplication, being divided across different paralogues.

In the case of *Hox* genes, I know that a complete subfunctionalization of the vertebrate *Hox* paralogues has not happened, as all paralogues still produce mRNAs and proteins of full-length *Hox* products. However, when I study the average rates of DRP across the *Hox* paralogue groups of *Homo sapiens*, I see that there is an asymmetrical distribution, with two peaks on PGs 3 and 6-9. Interestingly, I see that this heterogeneity is conserved across vertebrates: I also observe this asymmetry in *Mus musculus* and *Danio rerio*, and when I perform correlation analyses, I see that the profiles of DRP across PGs are indeed positively correlated. Moreover, when I compare the incidence of DRP across paralogues of the same organism, I see that the profiles are not significantly correlated, suggesting that in each organism, different *Hox* paralogues contribute mRNAs to the conserved differential RNA processing rates across the PGs of vertebrates.

I see another evolutionary dependence in our dataset: when the rates of *Hox* protein divergence within a paralogue group are compared with the respective average rates of DRP, I observe no relationship between the two measures for most PGs; however, posterior *Hox* genes of the *Abd-B* group (*Hox9-13*) have higher divergence rates than other PGs, and that these rates of protein evolution are inversely correlated with alternative mRNA production in a statistically significant way. As these loci are more recent than the remaining *Hox* genes of mammals, being the result of tandem duplications of an ancestral *Abd-B Hox* gene at the base of the chordate lineage, it is possible that I am capturing two kinds of relationships between alternative isoform production and protein evolution. I propose that for relatively recent genes, like the *Abd-B* class of mammalian *Hox* genes, loci undergoing fast evolutionary rates at the protein level have, at the same time, a constraint in the production of differential mRNAs; genes of the *Hox9-13* groups are not only similar to other paralogues, but to

genes in other PGs, whereas PGs *Hox1-8* have had mostly independent evolutionary lineages since the Urbilaterian. As such, I expect the selective pressure to change the protein-coding sequence to be more pronounced in *Hox9-13* groups. It is possible that in this scenario, protein evolution takes precedence, as the redundancy between a high number of genes is extraordinary; in cases when this is less pronounced (as with PGs *Hox1-8*), and paralogue loci are evolving at a slower rate, I would expect paralogues to accumulate alternative mRNAs in a manner that is dependent on other paralogues of the same PG, but less dependent on protein evolution. It would be interesting to explore this relationship in other duplicates in mammalian genomes, as mammalian *Hox* clusters specifically show a great degree of compaction, even when compared with *Hox* clusters of other chordates (Duboule 2007). It could be that this constraint in *Hox*-specific, as genomic compaction reduces the amount of sequence evolution that is tolerated in *Hox* clusters. For instance, the *Hox* clusters of Vertebrates display a clear lack repetitive elements (Fried et al. 2004). It is then possible that there is a strong constraint at the DNA sequence level in *Hox* clusters, with protein evolution taking precedence over the accumulation of alternative mRNA isoforms, also because some of the nucleotides in a *Hox* locus are involved in both open-reading frame sequences and alternative isoform production (see **Chapter 4**).

During the course of this work, I have studied the sequence of the *Ubx* transcriptional unit in 12 Drosophilids, looking at the distribution and conservation of targets for the *Drosophila melanogaster* neuronal-specific ELAV, an RNA-binding protein (Rogulja-Ortmann et al. 2014). In that work, I show that *Ubx* contains a host of putative sites for ELAV-mediated regulation of alternative RNA processing (Rogulja-Ortmann et al. 2014). A subset of these target sites was then experimentally shown by our co-authors to interact with ELAV *in vitro* and *in vivo*, a molecular partnership that

leads to the appropriate regulation of differential RNA processing in all *Hox* genes of the BX-C (Rogulja-Ortmann et al. 2014).

In this work, I look at the manner in which differential *Hox* mRNAs are produced in mammals, and find that this occurs by coordination of distinct kinds of transcription, alternative splicing and alternative cleavage and polyadenylation. Moreover, I see that there are two coordinated modes of alternative mRNA production, which integrate specific RNA processing events at the transcriptional and splicing levels. Interestingly, I also see that loci of the same PG tend the have similar differential RNA processing modes, indicating yet again that the list of similarities between paralogous *Hox*, which includes relative genomic context, sequence, expression and function, should also include differential RNA processing.

I have previously studied the evolution of miRNA targets in the 3'UTRs of the *Hox* gene *Ubx* (Patraquim et al. 2011), and shown, that the signals that mediate the formation of *Ubx* alternative 3'UTRs in *Drosophila melanogaster* are conserved across twelve Drosophilids, and that miRNA targeting by developmentally relevant miRNAs is predicted to remain within one of the two alternative 3'UTR isoforms of *Ubx* during the evolution of this gene (Patraquim et al. 2011). In this chapter, I study the generation of tandem 3'UTRs, the most overrepresented individual DRP event in our dataset, in the context of miRNA regulation. miRNAs have previously been shown to regulate *Hox* genes during development in both *Drosophila melanogaster* (Bender 2008; Thomsen et al. 2010) and *Mus musculus* (Hornstein et al. 2005). I find that APA generates alternative 3'UTR isoforms with different miRNA complements, as alternative 3'UTR tracts are more often than not anti-correlated in a significant way. However, when I compare orthologous 3'UTR isoforms across mammalian species, I find that proximal (or constitutive) 3'UTR sequences are conserved across mammals, and have similar

miRNA target complements, while distal 3'UTRs have significantly diverged in their miRNA target site complement. I propose that APA generates developmental compartments in the 3'UTRs of *Hox* genes. As distal 3'UTRs are usually deployed in the CNS of both *Drosophila* and mammals (Thomsen et al. 2010; Hilgers et al. 2012; Miura et al. 2013), I anticipate that novel 3'UTR-miRNA regulatory interactions, accumulated privately by either *Mus musculus* or *Homo sapiens*, should manifest themselves in this tissue. miRNA-based regulation is predicted to contribute to 2%–4% of mRNA and 4%–6% of protein expression differences across the brains of primates. If APA mediates differential visibility of products of a given locus to miRNA regulation (Thomsen et al. 2010) in the context of development and evolution, I expect that the transcriptome of mammalian central nervous tissues, both embryonic and adult, should reflect this regulatory level.

In this Chapter, I focus on protein-coding genes. However, *Hox* loci also display high levels of antisense transcription. For example, HOTAIR, a long noncoding antisense RNA that sits in the mammalian *Hox* clusters, has been shown to repress *HoxD* genes by enhancing repressive chromatin states (Li et al. 2013). This interaction leads to homeotic transformations in both the axial skeleton and the limb of *Mus musculus* (Li et al. 2013). It would be interesting to explore the relationship between the long noncoding RNA products of mammalian *Hox* clusters and the production of alternative protein-coding mRNAs, as *Hox* genes show high amounts of antisense transcription (Mainguy et al. 2007). Long noncoding RNAs haven been shown to regulate chromatin states in *Hox* loci (Li et al. 2013); chromatin states, in turn, have been shown to impact differential RNA processing (Acuña & Kornblihtt 2014; Luco et al. 2011). It remains to be seen whether the effects of long noncoding RNAs like HOTAIR extend to the regulation of differential RNA processing of *Hox* genes, as

HoxD genes, for instance, produce a host of alternative mRNA isoforms with functionally different protein domains (see **Chapter 4**).

Chapter IV

The production of Homeodomain-less Hox *isoforms by differential RNA processing*

N.B: Some experiments discussed in this Chapter include the contributions of others: the dissections of *Mus musculus* embryos and adult tissues were performed in collaboration with Claudio Alonso, Sofia Pinho and Aalia Bano. Additionally, Aalia Bano performed RNA extractions, RT-PCRs and Agarose Gel Electrophoresis using *Mus musculus* embryonic and adult tissue samples

4.1 – Chapter Overview

In the previous chapter I show that differential RNA processing significantly remodels the mRNA sequences of mammalian *Hox* genes. I note that some *Hox* paralogue groups produce more alternative isoforms than others and that these hotspots of mRNA remodeling within *Hox* clusters are conserved across vertebrates. Additionally, I show that the majority of *Hox* mRNA isoforms is formed through a series of coordinated regulatory steps that involve transcription, mRNA splicing and 3'UTR formation. Finally, I show that the constitutive 3'UTRs of *Hox* genes have significantly maintained their *cis*-regulatory complement across mammalian evolution, while the elective 3'UTR tracts significantly diverged between humans and mice, and propose that the production of alternative 3'UTR isoforms introduces compartments on both developmental and evolutionary scales.

In this chapter, I focus on the effects of alternative *Hox* mRNA processing on the production of alternative Hox protein isoforms. I report that at least eleven *Hox* genes produce mRNAs that do not encode for a Homeodomain (*Homeodomain-less*) in mammals, and that the resulting protein isoforms retain, in many cases, protein-protein interaction sequence modules. Additionally, I see that some *Hox* genes undergo differential RNA processing so as to produce Homeodomain-bearing isoforms that lack the (I)YPWM(K) hexapeptide (HX), a protein-protein interaction motif, as well as the SSYF transcriptional activation domain. I also observe that, in some cases, differential RNA processing introduces variation in the length of an amino acid stretch that lies between the HX and the Homedodomain – the Linker region. All of these effects of alternative RNA splicing on protein composition are expected to affect the molecular function of *Hox* proteins. Furthermore, I inspect PG10-specific protein motifs to show that alternative splicing is expected to impact Hox10 ability to repress rib formation in

the lumbar region in the development of the tetrapod axial skeleton. I hypothesise that this specific molecular function of Hox10 proteins has evolved before the emergence of rib repression in the axial skeleton, which in turn influenced the emergence of *Hoxa10* alternative splicing in mammals.

Finally, I develop an experimental approach to study the mechanisms leading to the production of Homeodomain-less isoforms, using the human *Hoxa9* gene in human Embryonic Kidney(HEK)-293 cell cultures as an ex-vivo model. I find that the production of the Homeodomain-less form occurs only after the HD+ has accumulated in human cells, and that this is likely not the result of recursive splicing as the production of Hoxa9 Homeodomain-less isoforms ceases once transcription is blocked. I then show that other 5 *Hox* genes (i.e. Hoxa1, Hoxb1, Hoxb9 and Hoxc4) also produce Homeodomain-less isoforms in human cells (experiments carried out by Aalia Bano in the Claudio Alonso Lab). Finally, I show that *Hox* genes produce Homeodomain-less isoforms during mouse embryogenesis and adulthood, and that the balance between HD+ and Homeodomain-less changes across time and space, suggesting that the production of Homeodomain-less isoforms is regulated *in vivo* (this work was performed by Dr. Claudio Alonso, Sofia Pinho and mostly Dr. Aalia Bano). I extend our observation to other mammalian Homeodomain *loci* outside the *Hox* family, and show that these too produce alternative mRNAs that do not encode for a Homeodomain in a conserved manner. I then use freely available transcriptome-wide data to explore the production of DNA-binding domain lacking (*DBD-less*) proteins in mammals. I show that other major Transcription-Factor classes, like zinc finger, leucine-zipper and bHLH proteins, also show evidence for the production of *DBD-less* isoforms, but that *Homeodomain-less* isoforms display the largest amount of conservation across mammals. I observe the production of *Homeodomain-less* isoforms in all metazoa

analysed, as well as in plants, and see many instances where homologous Homeodomain genes produce *Homeodomain-less* isoforms in both *C. elegans*/*Drosophila* and mammals. Our results collectively show that differential RNA processing significantly remodels the availability of functional protein motifs in Hox products, and suggest that further studies on Homeodomain genes require the inclusion of this aspect of gene regulation.

4.2 – Results

4.2.1 – Differential RNA processing produces *Hox* mRNAs that do not encode for the Homeodomain.

In this chapter, I investigate the consequences of mRNA processing on *Hox* protein domains. I have previously shown that the alternative processing of *Hox* mRNAs remodels open-reading frames and is thus expected to produce alternative protein isoforms from 64% of observed mRNA variants. Previous studies into the function of *Hox* proteins have uncovered a repertoire of protein modules that underlie important functions of this gene family. Among them, the DNA-binding Homeodomain stands out as the most remarkable *Hox* domain, as its sequence and function in mediating the transcriptional activation and repression of Hox targets are conserved across all Hox genes across metazoans. Additionally, the protein-protein interaction hexapeptide motif (YPWM/HX) has been shown to mediate the successful recruitment of PBC transcription factor proteins that work as Hox molecular partners by co-operatively binding to specific DNA targets in both arthropods and mammals. The latter motif is present in mammalian *Hox* Paralogue groups 1-8, while a more degenerate version containing a single conserved tryptophan amino acid (W) is present in *Hox9-13*. The

stretch of amino acids that stands between the HX and the Homeodomain (HD) is known as the linker region, and its length has been shown to act together with the HX to mediate the transcriptional activity of the *Drosophila Hox* gene *Abd-A*. Additionally, other Hox peptide motifs mediate general Hox function – the N-terminal SSYF, for instance, mediate the transcriptional activation of targets by Hox proteins. Other protein domains, like the Hoxa9 N-terminal activation domain, the M1 and M2 Hox10 domains and the Hoxa13 N-terminal domain, are specific to each vertebrate paralogue group, and are thus thought to have emerged in the ancestral *Hox* cluster at the base of the chordate lineage.

As all of the aforementioned Hox protein domains have specific, experimentally defined functions, I decided to investigate their occurrence as well as their possible exclusion in Hox proteins by means of differential RNA processing. To do so, I first used a biased approach, scanning all alternative, as well as reference Hox protein isoforms (see Chapter 3) for domains that are already annotated in online protein domain databases. More specifically I used the pre-computed InterProScan protein-domain predictions already present in the *Hox* isoform annotations of *Ensembl*. These predictions are biased, as they were performed by querying a given protein sequence for the existence of annotated domains, which are included in a variety of distinct protein-domain member databases like SMART, PIRSF, Pfscan, PRINTS and Pfam. This approach adds the value of redundancy to our analysis, as parallel databases have different and specific domain-identity requirements. This means that a given Hox domain (like the Homeodomain) will be identified by more than one prediction tool, which reduces the amount of false negative results in our study. This is specially important for our study, as one of the characteristics I am willing to explore is exactly the *exclusion* of specific *Hox* protein motifs in alternative protein isoforms.

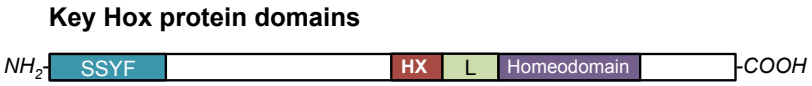
I first downloaded all InterProScan predictions for all Hox protein isoforms encoded by well-supported, GENCODE annotated protein-coding mRNAs (See Chapter 3). This yielded a list of 32 annotated domains that are shared by Hox proteins within both the human and mouse Hox protein complements. I then tabulated the presence or absence of each of the 32 InterProScan domains in each of the alternative Hox proteins in both humans and mice, and used this information to hierarchically cluster Hox proteins based on the their protein-motif complement. With this approach, I find that I can recapitulate the paralogue group (PG) membership of the overwhelming majority of Hox isoforms using the motifs available. This is true for both mice and humans. Interestingly, I see that Hox isoforms that do not group with their respective PG members are clustered together, in both mice and humans, in a cluster that is characterized by the absence of most motifs, including the Homeodomain. This result opened up the possibility that a host of Hox proteins (5 Hox *loci* in mice and 6 in humans) lack the Homeodomain. This result confirms that the differential processing of *Hox* mRNAs leads to changes in mRNA sequence predicted to significantly impact on the molecular function of Hox proteins.

Unfortunately, motifs like the Hexapeptide and the SSYF domain are absent from InterPro member databases and their inclusion/exclusion could not be assessed with this analysis. To further confirm the absence of Homeodomain motifs in some mammalian Hox protein isoforms, as well as to probe the inclusion and exclusion of other, aforementioned motifs of relevance for Hox function, I next performed an unbiased search for common protein domains in all mammalian Hox isoforms using the motif-search tool MEME. This tool finds novel, ungapped motifs in an unbiased manner, by comparing all sequences provided in parallel. In our case, I submitted the previously mentioned protein sequences for all Hox isoforms in both mice and humans,

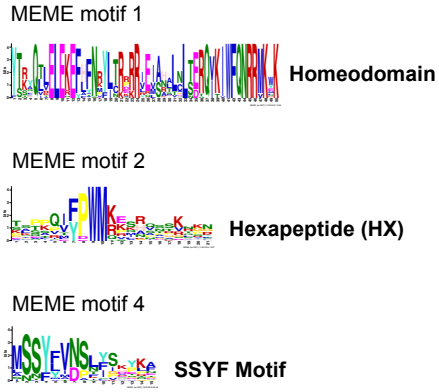
and chose to query this set of sequences for the 30 top motifs regardless of size (6-50 amino acids in length).

Our query recovered a list of 30 statistically significant protein motifs that are shared by 2 or more Hox protein isoforms. Upon initial inspection of the results, I found that the top ranked motif (Motif 1) in this list corresponds to the Homeodomain. Additionally, Motifs 2 and 4 respectively include the full sequences for both the Hexapeptide and the SSYF domains (**Figure 4.1**). I considered these results to be encouraging, in that, unlike the InterProScan predictions, our unbiased method is reliably able to assess the presence and absence of important Hox domains like the HX and the SSYF. As such, I repeated the hierarchical clustering analysis of Hox proteins based on the presence/absence of protein domains, only this time using the results of our MEME query. I find that, as with the previous analysis, most Hox isoforms group with other isoforms of the same paralogue group. Interestingly, I confirm that thirteen Hox isoforms lack the Homeodomain: *Hoxa1*, *Hoxa9*, *Hoxa10*, *Hoxb1*, *Hoxb3*, *Hoxc11* and *Hoxd12* in humans, *Hoxa1*, *Hoxa9*, *Hoxa7* (two different isoforms), *Hoxb9* and *Hoxc4* in mice (**Figures 4.1C-D** and **Figure 4.2**). This indicates that a variety of *Hox* genes produce mRNA isoforms that do not encode for a Homeodomain by differential RNA processing, and suggests that the production of DNA-binding domain-less Hox proteins by DRP is a common theme in mammalian Hox genes. Although the production of Homeodomain-less *Hox* isoforms has been shown for *Hoxa1*, *Hoxa9* and *Hoxb6* in mammals, as well as *Hoxb7* in *Xenopus laevis* embryos (Fernandez & Gudas 2009; Shen et al. 1991; Hong et al. 1995; Fujimoto et al. 1998; Wright et al. 1987), I do not see *Hoxb6-Homeodomain-less* isoforms in our analysis. However, our observations significantly expand the repertoire of *Hox* genes that undergoes this differential RNA processing mechanism. Additionally, I see that most *Homeodomain-*

A



B



C

Column1	SSYF Motif	Hexapeptide/*W*	Linker Variation	Homeodomain
H-Hoxa1-002	*NA*	0	0	0
H-Hoxa10-201	✔	0	0	0
H-Hoxa9-003	*NA*	0	0	0
H-HOXB1-002	*NA*	✔	0	0
H-HOXB3-006	*NA*	0	0	✔
H-HOXB3-008	*NA*	0	0	✔
H-Hoxb3-009	*NA*	0	0	0
H-Hoxc11-002	*NA*	0	0	0
H-Hoxd12-002	*NA*	0	0	0
H-HOXD8-002	0	0	0	✔
M-Hoxa1-002	*NA*	0	0	0
M-Hoxa7-003	✔	✔	0	0
M-Hoxa7-004	0	0	0	0
M-Hoxa9-002	*NA*	0	0	0
M-Hoxb9-002	✔	0	0	0
M-Hoxc4-001	✔	✔	0	0
M-Hoxd8-003	0	0	0	✔

D

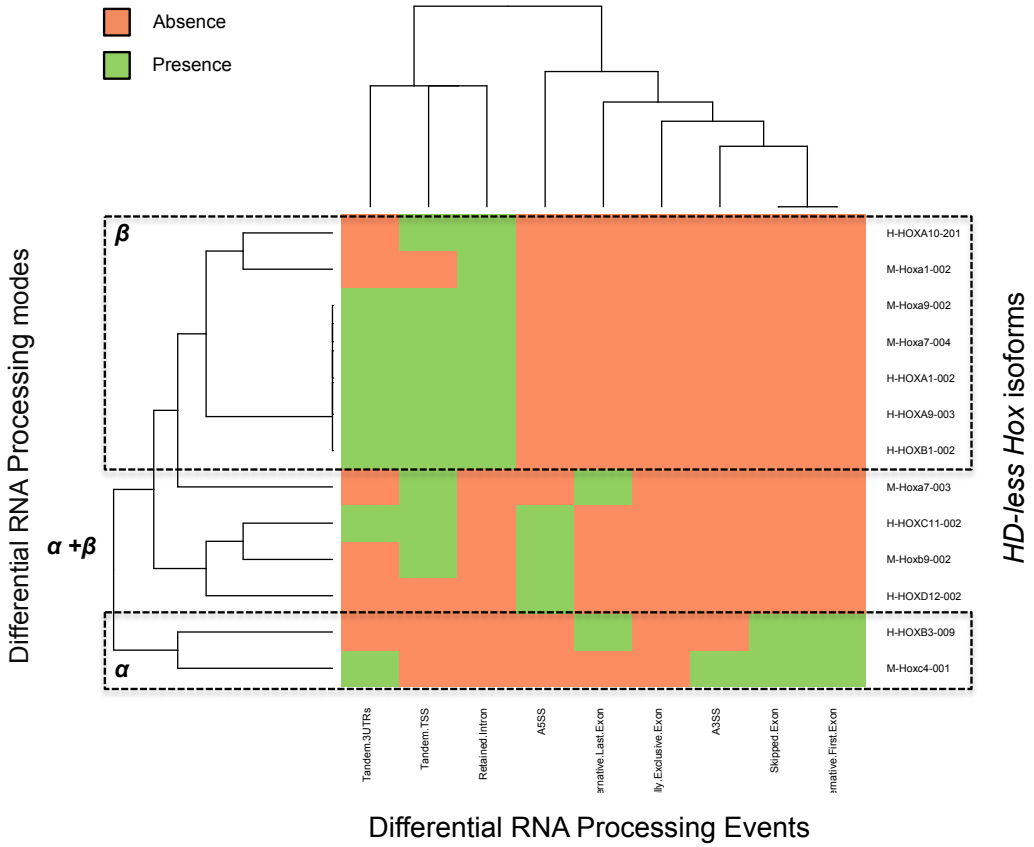


Figure 4.1 – An unbiased search for Hox peptide motifs recovers key Hox domains involved in the molecular function of Hox proteins (legend in the following page)**.**

Figure 4.1 – An unbiased search for Hox peptide motifs recovers key Hox domains involved in the molecular function of Hox proteins. (A) Key Hox protein domains. Most mammalian Hox proteins contain the SSYF, hexapeptide (HX) and the Homeodomain. The amino acid sequence between the hexapeptide and the Homeodomain is usually called the Linker region (L). The SSYF motif mediates the activation of transcription by Hox transcription factors; The HX mediates the interaction between Hox and PBC proteins (see **Figure 1.3B**). The Homeodomain mediates the interaction between Hox transcription factors and their DNA targets. **(B)** An unbiased search for Hox protein motifs recovers key Hox protein domains. I used MEME (Bailey et al. 2009) as an unbiased method to inquire about the inclusion and/or exclusion of functional Hox domains (see panel A) in alternative Hox proteins. Our approach recovers key Hox protein domains, like the Homeodomain (MEME Motif 1), the HX (MEME motif 2) and the SSYF motif (MEME Motif 4). The number of each MEME motif represents its rank in the analysis, with the Homeodomain (Motif 1) being the most common. **(C)** The differential RNA processing of *Hox* mRNAs introduces combinatorial variation in the form of presence/absence of key Hox protein domains. Strikingly, I observe that there are thirteen isoforms in which the Homeodomain is not included in the final Hox protein. In some cases, like with the *Mus musculus* Hoxc4-001 isoform, Homeodomain-less isoforms contain other key Hox domains like the SSYF motif and the HX. **(D)** Distinct RNA processing modes are involved in the production of Homeodomain-less isoforms in mammalian *Hox* genes. The hierarchical clustering of Homeodomain-less encoding *Hox* mRNAs was compared to the differential RNA processing events involved in their production (See Chapter 3, **Figure 3.4A**). The two differential modes α and β are involved in the production of alternative *Hox* mRNAs that do not encode for the Homeodomain, indicating that different kinds of coordinated RNA processing modes can lead to similar outcomes at the level of Hox protein-sequences.

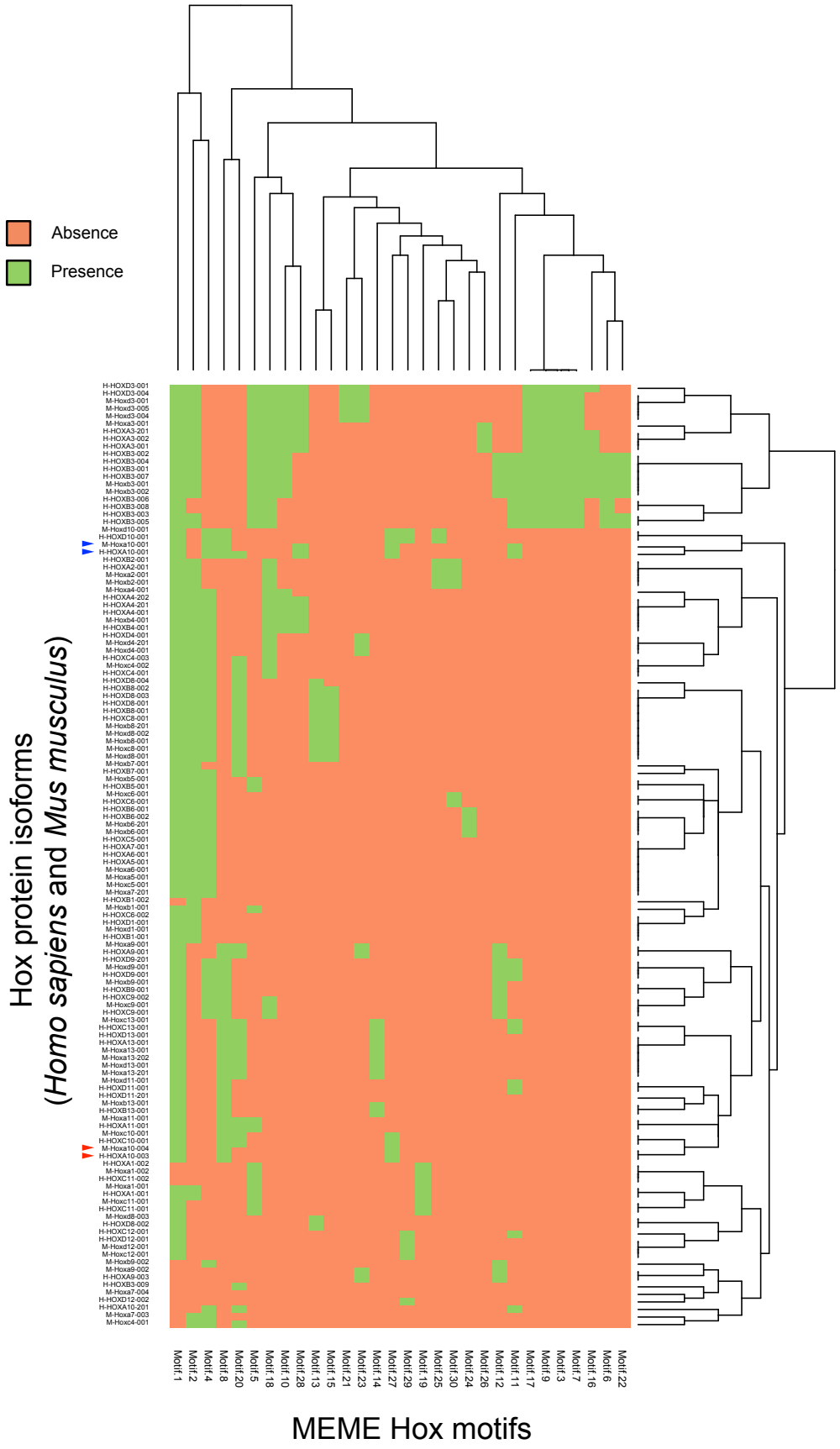


Figure 4.2 – Hierarchical clustering of Hox peptide motifs groups alternative Hox isoforms of the same paralogue group (legend in the following page)**.**

Figure 4.2 – Hierarchical clustering of Hox protein motifs groups alternative Hox isoforms of the same paralogue group. Hierarchical clustering analysis of MEME results after an unbiased search for Hox protein motifs in all alternative Hox proteins of *Mus musculus* and *Homo sapiens*. I performed a hierarchical clustering analysis of all alternative Hox proteins in mammals, with the goal of understanding the variation introduced in Hox proteins by differential RNA processing. I observe that most Hox isoforms tend to cluster with other isoforms of the same PG. However, I see that two alternative Hoxa10 isoforms, HOXA10-003 and Hoxa10-004 in *H. sapiens* and *M musculus*, respectively, (red arrowheads) are more similar in sequence to products of the *Hoxc10* locus than to other proteins of the Hoxa10 locus (blue arrowheads). This result suggests that differential RNA processing can generate enough variation in Hox protein sequences to remodel the identity of Hox proteins. Note that these alternative Hoxa10 proteins contain a Homeodomain (Motif 1). Additionally, I see that most Homeodomain-less proteins cluster together. In most cases, these Hox isoforms contain sequence motifs that are shared with other Homeodomain-containing Hox proteins, suggesting that Homeodomain-less isoforms could perform some Hox-like molecular functions.

less isoforms of *Hox* genes are produced by a mix of tandem transcriptional initiation, followed by intron retention and tandem alternative polyadenylation (**Figure 4.1D**), placing them in the β mode of *Hox* mRNA processing (See **Chapter 3**).

Next, I studied the inclusion/exclusion by differential RNA processing of other protein domains of known importance for Hox protein function. I observe that four Hox isoforms lack the YPWM motif but show the presence of the Homeodomain, as is the case with *Hoxd8* in both humans and mice, as well as two *Hoxb3* isoforms in humans. This suggests the interesting possibility that, in the aforementioned cases, differential RNA processing leads to Hox proteins that can bind DNA, but have decreased ability to interact with the major Hox8 molecular partners – PBC proteins. I also find that ten Hox protein isoforms lack both the Homeodomain and the YPWM motif (or the degenerate NNWN motif, in the cases of Hox9-13). However, these proteins are not featureless, as they share a number of motifs with "full-length" Hox isoforms, including the SSYF activation domain in at least in four cases (**Figure 4.1C**).

Next, I wondered whether the Hexapeptide/Homeodomain Linker region was, in any case, remodeled by differential RNA processing. I see that, in one case – *Hoxd8* – Differential RNA processing reduces the size of the linker region from 6 to 5 amino acids. Additionally, I see that *Hoxb8* shows an identical remodeling of protein isoforms in humans (**Figure 4.1C**). These results are interesting, as the linker region of abd-A, the *Drosophila melanogaster* Hox homologue of mammalian Hox8 proteins, has been shown to act in conjunction with the Hexapeptide to promote an activation/repression switch in abd-A activity. In this specific case, a short peptide (PFER) within the Linker region, and not linker size *per se*, seems to be responsible for the postulated action of this region in concurrence with the Hexapeptide. Upon inspection, I have not identified any homologous PFER peptide in mammalian Hoxb8/Hoxd8 proteins. Rather, the use

of an alternative 3' splice site in the processing of Hoxd8 mRNAs seems to delete an Alanine residue (A) from the linker region. As such, a putative role of this linker size variation remains unclear for *Hox8* genes.

Finally, I observe that the SSYF transcriptional activation domain seems to be included or excluded in Hox proteins, as a result of differential RNA processing. In our analysis, I find that genes of the paralogue groups Hox4-10 show the native presence of this domain (with the exceptions of Hoxa9 in both mice and humans, as well as Hoxb7 in mice). However, I see that in four cases (Hoxc4, Hoxa7 and Hoxb9 in mice, Hoxa10 in humans), alternative Hox proteins that do not have the Homeodomain show the presence of the SSYF domain (**Figure 4.1C**). This result indicates that the aforementioned Homeodomain-less proteins, two of which also include the Hexapeptide domain, have functional domains despite lacking the Homeodomain. Additionally, two alternative Hoxd8 isoforms that lack the homeodomain also lack the SSYF motifs. This is also seen in one Hoxa7 isoform in mice. The latter is the only *domain-less* isoform in our analysis, and this suggests that none of its amino acid sequence is recognized by our analysis as having commonalities with the Hox families. However, for the remaining 12 Homeodomain-lacking isoforms, I see a complex, combinatorial exclusion/inclusion of other Hox functional domains.

Together, these results indicate that the differential RNA processing of *Hox* mRNAs significantly remodels the *anatomy* of Hox proteins, as features that have been shown to mediate the main aspects of molecular role of Hox proteins, like the SSYF domain and the hexapeptide, seem to be combinatorially included/excluded or tinkered with by differential RNA processing. Additionally, I show that some Homeodomain-lacking isoforms have Hox-specific functional domains, like the protein-interaction domain YPWM motif. These results indicate that Homeodomain-less isoforms are not

expected to completely lack function, and suggest that this mechanism plays an active role on Hox molecular function.

4.2.2 – M1 and M2 motifs in Hox10 proteins: a case study in the evolution and alternative splicing of functionally important Hox protein motifs.

Upon close inspection of the hierarchichal clustering analysis for *Hox* protein motifs (**Figure 4.2**), I noticed that two *Hoxa10* isoforms, *Mus musculus Hoxa10-004* and *Homo sapiens Hoxa10-003* were grouped together with *Hoxc10* isoforms, rather that with other isoforms of the *Hoxa10* locus. This indicates that differential RNA processing in the *Hoxa10* locus generates two alternative isoforms with protein motif complements that are closer to isoforms of another paralogous locus. This grouping is associated with the depletion of motifs in both *M-Hoxa10-004* and *H-Hoxa10-003*, when compared with other *Hoxa10* isoforms (**Figure 4.2)**.

Hox10 paralogues are expressed along the A-P axis of developing mammals in developing lumbar regions of the vertebral column. These genes show a great degree of functional redundancy in this context, acting in coordination to mediate the repression of rib formation (a thoracic fate) in lumbar vertebrae (Wellik & Capecchi 2003). The repression of ribs during lumbar vertebral specification is a derived character in vertebrates, while the posterior extension of ribs along the A-P axis represents the vertebrate "ground-state" (Wellik & Capecchi 2003).

In a recent study, the N-terminal region of *Hox10* genes, have been implicated, along with the paralogue group-specific *Hox10* motif M1, in the repression of rib formation in mammalian vertebrae(Guerreiro et al. 2012). The M1 motif lies

immediately before the homeodomain and shows conservation across tetrapods, as does motif M2, which lies immediately after the homeodomain sequence (Guerreiro et al. 2012), (**Figure 4.3**). M1 contains two amino acids, Serine and Threonine, both putative targets for phosphorylation, which when mutated into Alanine lead to a loss of rib-repressing activity, indicating that not only the motif sequence but its phosphorylation state might underlie the repressive ability of Hox10 proteins (Guerreiro et al. 2012).

I first wondered whether alternative splicing remodeled the availability of these motifs in protein products of the *Hox10* paralogue group. Both (Guerreiro et al. 2012) and (Benson et al. 1995) report a *Hoxa10* isoform that lacks the N-terminal region but includes the M1 and M2 motifs, as well as the homeodomain. I observe the occurrence of this isoform in our analysis, in both *Mus musculus* and *Homo sapiens*. To inquire about the effect of differential RNA processing in the availability of Hox10 protein motifs, I performed an unbiased MEME protein motif search, followed by a hierarchical clustering analysis of shared protein motifs. In addition to all differential Hox10 protein isoforms in both *Mus musculus* and *Homo sapiens*, I included the protein sequences of all Hox10 loci in the amphibian *Xenopus tropicalis*, the zebrafish *Danio rerio*, the cephalochordate *Amphioxus*, and the tunicates *Ciona intestinalis* and *Oikopleura dioica* (**Figure 4.3A)**. First I find that an unbiased search for protein motifs recovers 16 motifs. I hierarchically clustered these results, using the presence or absence of motifs as a clustering character (**Figure 4.3D)**. Our unbiased strategy successfully recovers a number of N-terminal motifs and the M1 and M2 domains, in addition to the Homeodomain (**Figure 4.3F**). Second, I find that in *Homo sapiens*, the *Hoxa10* locus produces a second alternative mRNA that lacks the

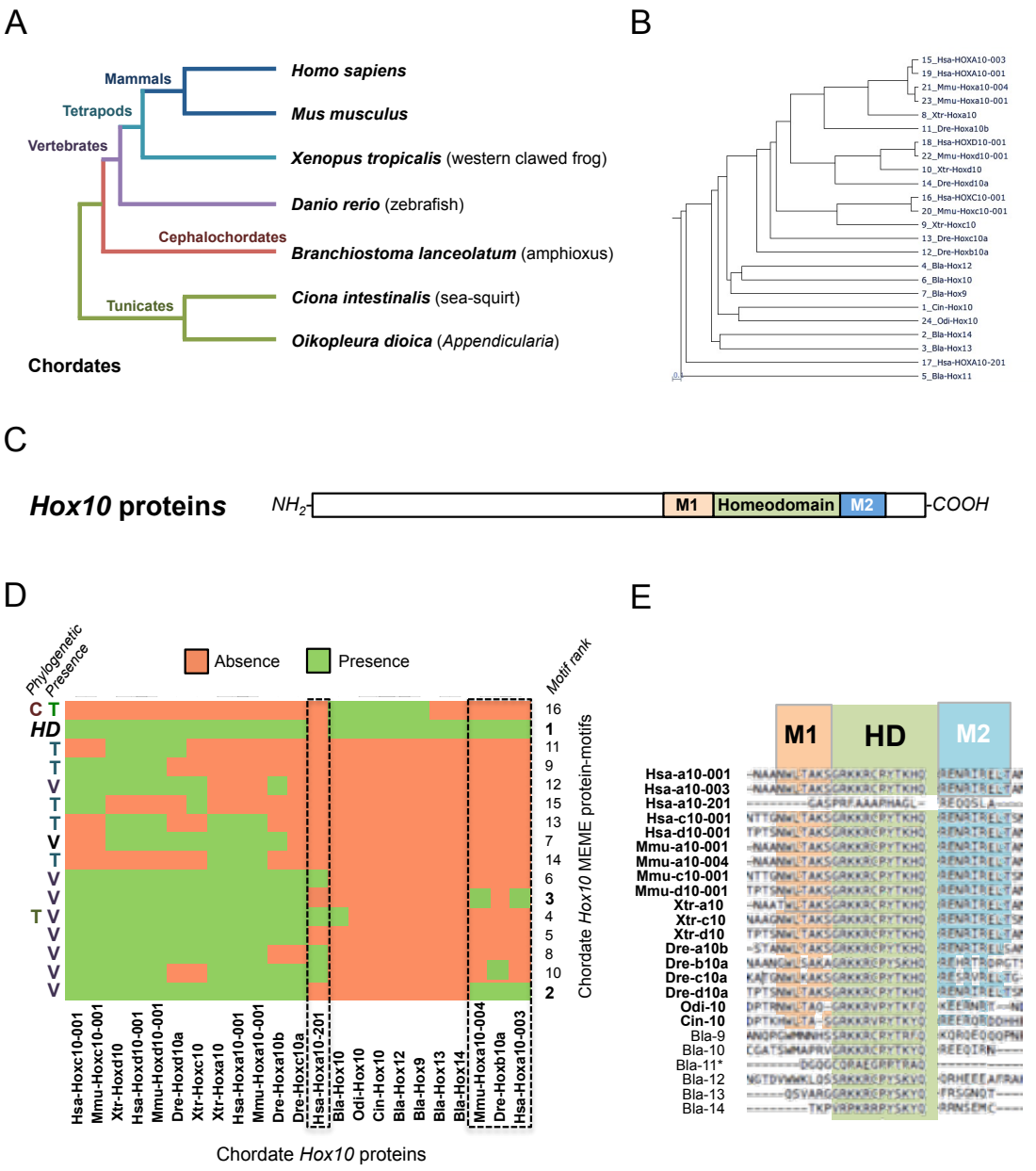


Figure 4.3 – Alternative splicing of *Hoxa10* generates atavistic Hox10 protein isoforms (legend in the following page)**.**

Figure 4.3 – Alternative splicing of *Hoxa10* generates atavistic Hox10 protein isoforms. (A) Classical phylogenetic relationships among the main clades of the chordate phylum, as well as the organisms used in this analysis. The phylum *Chordata* includes tunicates and cephalochordates, which are sister-taxa to the Vertebrates. **(B)** Cladogram of the phylogenetic relationships between chordate Hox10 proteins. A Neighbour-Joining phylogeny of chordate Hox10 proteins recovers relationships of orthology in most cases. However, the human HOXA10-201 isoform is not grouped within the Vertebrate Hox10 homologues. The protein sequences of all *Hox9-14* genes were used in the case of *Branchiostoma lanceolatum*, as the relationships of orthology between these genes and the vertebrate *Hoxa10* paralogues are unclear. **(C)** Key motifs in the Hox10 paralogue group in mammals. **(D)** Hierarchical clustering of Hox10 protein motifs. The alternative Hoxa10 isoforms 003 and 004 of humans and mice, respectively, contain protein motifs that group them with the zebrafish Hoxb10a orthologue. Conversely, the motifs in the human HOXA10-201 isoform group it with other vertebrate Hox10. This result stands in contrast with the phylogenetic relationships showed in panel B, and indicates that not all amino acids in Hox10 proteins are phylogenetically informative. **(E)** Conservation of Hox10 protein motifs across the Hox10 homologues of chordates. M1 and the M2 motifs are conserved beyond the tetrapod clade. The M2 motif is vertebrate-specific, while the M1 motif is present in full in at least two Hox10 orthologues of *Danio rerio*, with a slightly degenerate version in Urochordates. These results indicate that the M1 motif was co-opted for its rib-repressing developmental role during the evolution of tetrapods. **(F)** Main motifs recovered in an unbiased MEME analysis of chordate Hox10 protein-sequences.

M1 and M2 motifs, as well as the Homeodomain, but includes a number of other Hox10 motifs that are present in the N-terminal region of the full-length mammalian *Hox10* proteins, as well as in all vertebrate homologs (**Figure 4.3D)**. This result shows that alternative splicing can generate *Hox10* mRNAs that do not include the M1 and M2 motifs or the Homeodomain, but contain motifs in the rib-repressing N-terminal region. Further, it suggests that *Homo sapiens* can generate a version of a Hox10 protein that has the ability to repress rib fates in lumbar regions without directly binding DNA targets. Third, I find that the M1 motif (MEME motif 2 – see **Figure 4.3F**) is present in three of the four zebrafish *Hox10* paralogues (**Figure 4.3D)**. I decided to inspect this result more closely by analyzing the amino acid alignments of the Hox10 proteins analysed, and find that two zebrafish homologues, *Dre-Hoxa10b* and *Dre-Hoxa10b,* contain the full sequences for both M1 and M2 (**Figure 4.3E**), while two others, *Dre-Hoxc10a* and *Dre-Hoxb10a*, contain a slightly degenerate version of the M1 motif and a highly degenerate M2 sequence. Interestingly, the M1 sequence of the both isoforms has degenerated precisely on the aforementioned phosphorylation sites (**Figure 4.3E**): Finally, I find that tunicates but not cephalochordates, include a highly conserved version of the M1 motif; although this aspect has escaped our unbiased MEME search, I see the conservation of six out of seven amino acids in the M1 motif of *Oikopleura dioica* (including the Threonine phosphorylation site), and five out of seven in *Ciona intestinalis*, including both Serine and Threonine phosphorylation sites.

Our results suggest that the M1 motif has appeared in *Hox10* sequences early in chordate evolution, before the emergence of lumbar regions of tetrapods, or indeed before the appearance of the Vertebrates altogether. Conversely, the M2 motif, as well as the N-terminal regions, are exclusive to the vertebrate lineage. As such, our analysis indicates that the rib-repressing function of the tetrapod M1 motif was co-opted from an

earlier, unknown basal chordate function. In *Homo sapiens* and *Mus musculus*, differential RNA processing generates isoforms that group with the zebrafish *Hoxb10a* paralogue, in that they lack N-terminal regions. This indicates that differential RNA processing can generate an atavistic mRNA isoform, perhaps maintaining a pleiotropic function in old developmental contexts, basal to the chordates, as well as a novel rib-repressing function. Additionally, I see that another *Homo sapiens Hoxa10* isoform lacks M1, M2 and the Homeodomain, but includes Vertebrate-specific N-terminal-regions. As such, differential RNA processing of mammalian *Hox10* genes changes the availability of rib-repressing protein motifs in at least two opposite ways.

4.2.3 – The production of *Hox* mRNAs that do not encode for the Homeodomain is regulated in time and space during *Mus musculus* embryogenesis and adulthood.

In previous sections, I explored the generation of Hox protein isoforms that lack key protein domains, like the HX and the Homeodomain. To better understand the production of Homeodomain-lacking *Hox* isoforms in mammals, we first designed primers for all isoforms of *Hoxa1*, *Hoxa9*, *Hoxb1*, *Hoxb9* and *Hoxc4* mRNAs, in both mice and humans (**Figure 4.4**). We chose to analyse *Hoxa1* and *Hoxa9* as these show the conserved production of *Homeodomain-less* mRNAs (**Figure 4.2**). In the case of *Hoxb1*, this gene produces a *Homeodomain-less* isoform in humans, and we wondered whether this was also true in mice, despite the lack of evidence for this differential RNA processing event in the GENCODE annotation. The same is true for *Hoxb9* and *Hoxc4* in mice, where the existence of a *Homeodomain-less* isoform in humans is not observed (**Figure 4.4**).

Next, we isolated *Mus musculus* embryos at five developmental stages (8.5, 9.5, 10.5 11.5 and 12.5 d.p.c.). This was followed by the RNA isolation and subsequent RT-PCR analysis for each of the five aforementioned genes in all five developmental stages. In all *Hox* genes analysed, we find that both a Homeodomain-containing and a Homeodomain-lacking mRNA isoform is observed during embryogenesis (**Figure 4.4**). This is the first report of a *Hoxb1 Homeodomain-less* mRNA in *Mus musculus*. Moreover, all *Hox* genes show expression at all developmental stages analysed (**Figure 4.4**). In the case of *Hoxa1*, we see that the Homeodomain-containing isoform is the preponderant one across development, with the *Homeodomain-less* version of this gene being always present at lower concentrations, with an expression peak at 10.5 d.p.c. (**Figure 4.4**). The *HD-containing* isoform is also preponderant in the case of *Hoxb9*, being expressed at high levels in all developmental stages. In this case, the *Homeodomain-less* isoform only appears at 9.5 d.p.c. at low levels, increasing in expression between 10.5-12.5 d.p.c. (**Figure 4.4**). In the case of *Hoxa9*, both isoforms are present in all stages, and at high concentrations. Conversely, we find that for *Hoxc4*, The *Homeodomain-less* isoform is highly expressed in all stages, with the *Hoxa4-full* isoform being expressed in all stages, albeit at lower levels (**Figure 4.4**). Finally, the case of *Hoxb1*, we discover the existence of a Homeodomain-lacking isoform in *Mus* musculus, finding, surprisingly, that this *Homeodomain-less* isoform is preponderant in all stages analysed, and specially enriched in 8.5-10.5 d.p.c., while the *Hoxb1-full* isoform peaks in expression at 9.5-10.5 d.p.c (**Figure 4.4**).

Finally, Dr. Aalia Bano has recently observed the production of *Homeodomain-less* isoforms for all five aforementioned *Hox* genes in human HEK293 cells. This indicates that the production of *Hoxc4* and *Hoxb9 Homeodomain-less* isoforms is conserved between humans and mice, and significantly expands the

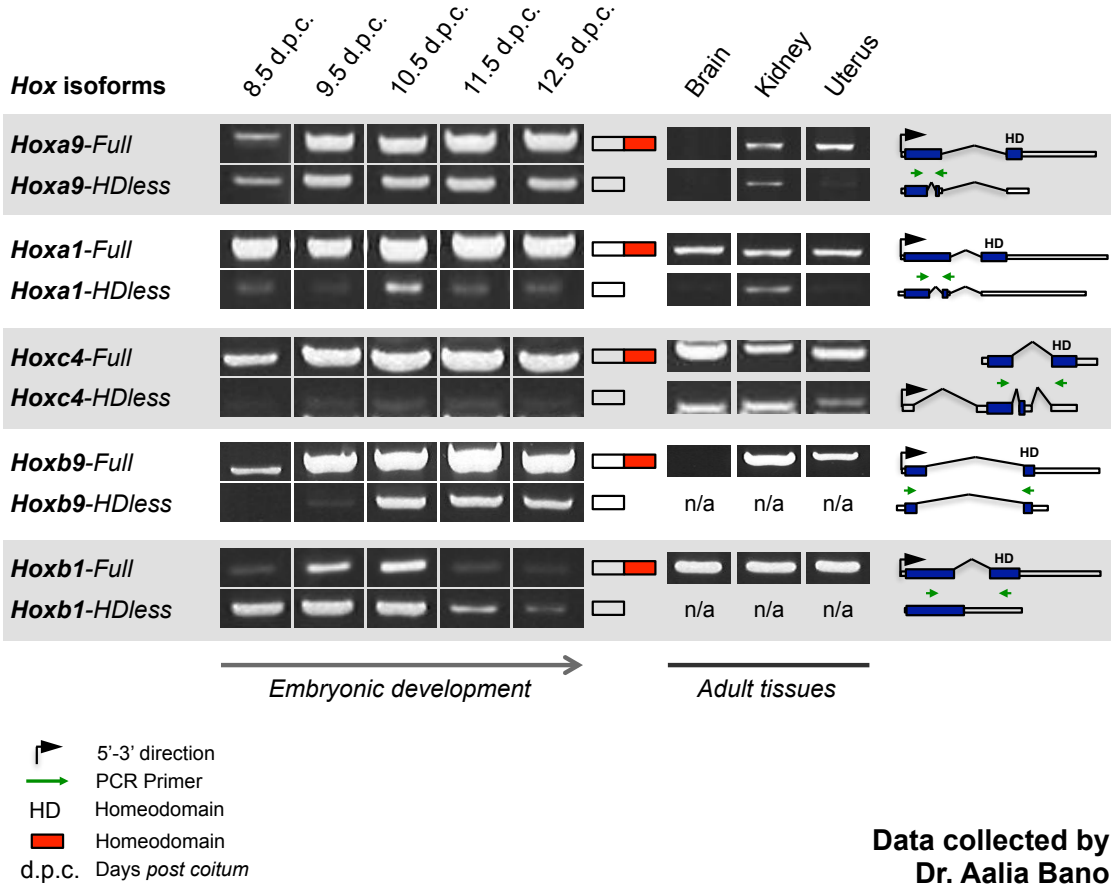


Figure 4.4 – *Mus musculus Hox* genes *Hoxa1, Hoxa9, Hoxc4, Hoxb9* and *Hoxb1* produce mRNAs that do not encode for a Homeodomain in a developmentally regulated manner (legend in the following page)**.**

Figure 4.4 – *Mus musculus Hox* genes *Hoxa1, Hoxa9, Hoxc4, Hoxb9* and *Hoxb1* produce mRNAs that do not encode for a Homeodomain in a developmentally regulated manner. Molecular analysis of Homeodomain-less expression during the embryonic development and adulthood of *Mus musculus***.** The *Hoxa1, Hoxa9, Hoxc4, Hoxb9* and *Hoxb1* loci produce alternative mRNAs that do not encode for the Homeodomain (see **Figures 4.1** and **4.2**). We observe the production of Homeodomain-less isoforms for all the aforementioned loci during the development of *Mus musculus*. With the exception of *Hoxc4*, all other loci produce large amounts of Homeodomain-less isoforms in a developmentally regulated manner. *Hoxa9* produces Homeodomain-less isoforms throughout the developmental window analysed, while *Hoxa1* and *Hoxb9* show an increase in Homedomain-less production at 10.5 d.p.c. Conversely, *Hoxb1* shows higher levels of Homeodomain-less isoforms at earlier stages (8.5-10.5 d.p.c.). In adult tissues, the production of Homeodomain-less isoforms also shows signs of specific regulation. Interestingly, the Homeodomain-less isoforms seem to be produced at similar rates as the Homeodomain-containing versions of *Hoxa9*, *Hoxa1* and *Hoxc4* in the adult Kidney. Conversely, the Homeodomain-less mRNAs from these loci seem to be mostly absent in the Brain and Uterus, with the exception of *Hoxc4*. These results show that *Mus musculus* produces isoforms that do not encode for the Homeodomain in a regulated manner across both developmental time and adult tissues.

documented cases of evolutionary conservation of this regulatory outcome in mammalian *Hox* gene expression.

Together, these results indicate that at least five of *Hox* genes express mRNA isoforms that do not encode for the Homeodomain in human cultured cells and throughout the development of *Mus musculus*. In mice, I see that this process is regulated in time, with the balance between Homeodomain-containing and Homeodomain lacking isoforms being different for different *Hox* genes. Interestingly, most of the signal corresponds to the *Homeodomain-less* in the cases of *Hoxc4* and *Hoxb1*, indicating that this isoform could have an important function in early mammalian development.

To further understand the expression of *Homeodomain-less* isoforms *in vivo*, I dissected the Brain, Kidney and Uterine organs of adult female mice, and asked whether *Homeodomain-less* isoforms were expressed in wild-type adult mice (**Figure 4.4**). Using the same primers as in the aforementioned analysis, I find that *Hoxa9* has low to no expression in the adult brain, but shows the expression of both *Hoxa9-full* and *Hoxa9-Homeodomain-less* isoforms in both Kidney and Uterus (**Figure 4.4**). In the Kidney, there is a balance between these two isoforms, whereas in the case of the Uterus, the Homeodomain-containing isoform is more preponderant. For *Hoxa1*, I find that the Homeodomain-containing isoform is the main mRNA from this locus in all tissues, but that the *Homeodomain-less* isoform is expressed at low levels in all three biological contexts (**Figure 4.4**). In the case of *Hoxc4*, I find that the *Homeodomain-less* isoform, as with embryonic development, is preponderant in the adult Brain and Uterus, and has a similar, high expression to the *Hoxc4-full* isoform in the adult Kidney (**Figure 4.4**). For *Hoxb1* and *Hoxb9*, I can only report the expression of the Homeodomain-containing isoforms in adult tissues (with the exception of *Hoxb9* in the

brain, where no expression is seen at all) (**Figure 4.4**). Additionally, Dr. Aalia Bano has since confirmed the expression of both *Hoxa1* and *Hoa9 Homeodomain-less* isoforms at the protein level (data not shown) in both embryogenesis and adult tissues, by Western Blot. This indicates that these mRNAs are translated, as annotated by GENCODE, and that the expression patterns at the mRNA level approximate the *Hox* isoform protein expression level.

Together, these results confirm the existence of *Hox* mRNA isoforms that do not encode for the Homeodomain in both embryonic development and adult physiology, and show that the *in vivo* balance between Homeodomain-containing and Homeodomain-lacking isoforms changes in time and space in *Mus musculus*. These results further solidify the notion that the production of Homeodomain-lacking isoforms is widespread in mammalian *Hox* genes, and points to an important role of differential RNA processing in the regulation of *Hox* gene expression in embryonic development and adult physiology.

4.2.4 – The human gene *Hoxa9* produces mRNAs that lack the Homeodomain.

In order to understand the differential RNA processing mechanism by which Homeodomain-less Hox proteins are formed, I decided to study the production of mRNA isoforms that do not encode for the Homeodomain from the human *Hoxa9* gene.

I chose to focus on this gene, as *Hoxa9* shows the conserved production of Homeodomain-less forms in both mice and humans (**Figure 4.5**). Additionally, the Homeodomain-lacking (Homeodomain-less) *Hoxa9* isoform has been shown by others to be conserved between mammals and birds. The same authors show that this

Homeodomain-less Hoxa9 transcript seems to be expressed throughout embryogenesis, being specially enriched in the developing genital tract, kidney, forelimb and tail of *Mus musculus* (Dintilhac et al. 2004). These data suggest a tissue specific effect of Hoxa9 isoforms that lack a Homeodomain.

Furthermore, the alternative *Hoxa9* isoform in question does not encode for a Hexapeptide (an AAN**W**LH sequence in the case of Hoxa9), a peptide that has been shown to mediate Hoxa9 protein-protein interactions with Pbx1a. However, Shen and colleagues (Shen et al. 1996) have shown that a mutation in the central tryptophan amino acid of this domain (W) had no negative effect on the ability of Hoxa9 to bind to Meis1 and stabilize this heterodimer's interactions with DNA. Even though the Hoxa9 N-terminal region alone does not stabilize Meis1 binding to DNA, as mutations that delete the Hoxa9 carboxy-terminal Homeodomain but maintain the 204 N-terminal amino acids abolish the DNA-binding ability of Hoxa9 and its ability to stabilize Meis1 DNA binding. The authors of this study then show that the ability of Hoxa9 to bind Meis1 resides in the first 61 amino acids (deemed the MEIS interaction motif or "MIM"*)*, but that for this heterodimer to be stably bound to DNA, both N-terminal and Homeodomain regions need to be present in the Hoxa9 molecular partner.

These results are relevant for ascribing a functional role to the Homeodomain-lacking Hoxa9 form (Hoxa9-Homeodomain-less), as this protein isoform includes the first 105 amino acids of the full Hoxa9 (Hoxa9-Full) isoform, and thus the MIM, as well as a novel S amino acid added just before the STOP codon. As such, this data indicates that the Hoxa9-Homeodomain-less protein does not stabilize Meis1 interactions with the DNA, but that it can bind Meis1. In this scenario, Hoxa9-Homeodomain-less could function as a dominant negative. Furthermore, this isoform has been shown to mediate the leukaemogenic function of Hoxa9 in Acute Myeloid

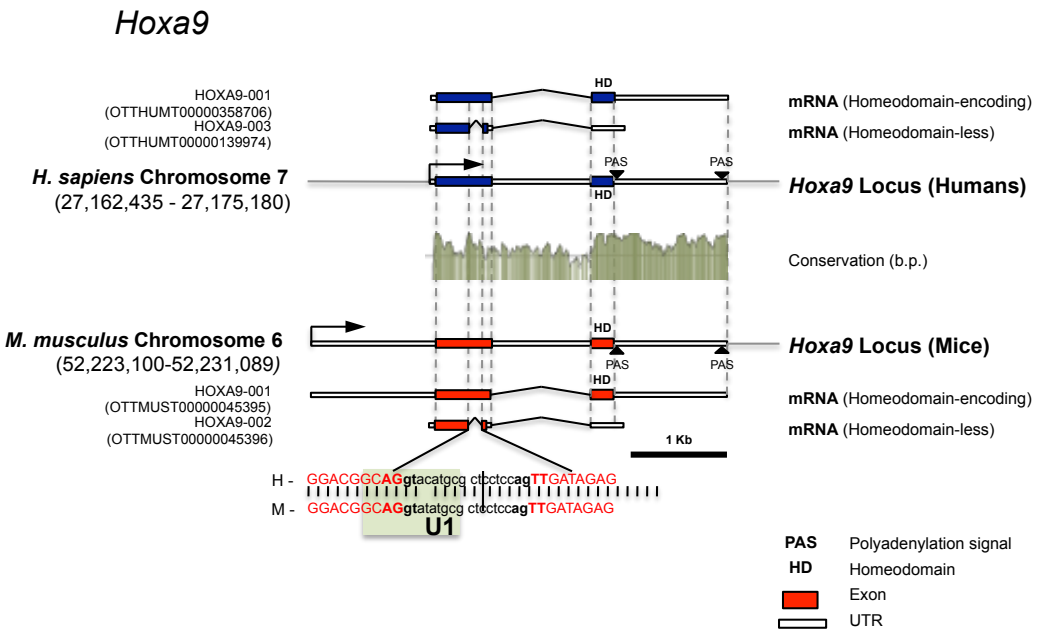


Figure 4.5 – Alternative splicing of *Hoxa9* produces Homedomain-encoding and Homeodomain-less mRNA isoforms in mammals (legend in the following page)**.**

Figure 4.5 – Alternative splicing of *Hoxa9* produces Homedomain-encoding and Homeodomain-less mRNA isoforms in mammals. The *Hoxa9* locus produces alternative mRNAs that do not encode for the homeodomain in a conserved manner. In both mice and humans, the GENCODE-annotated alternative *Hoxa9* mRNAs are produced, first, by the excision of an intronic sequence, which leads to a frameshift in the *Hoxa9* open-reading frame and a consequent premature STOP codon. However, transcriptional elongation continues across the locus, and a second constitutive intron is excised. The Homeodomain-encoding Homeobox sequence, lying downstream of these processing events, becomes included in the 3'UTR of alternative *Hoxa9* mRNAs. The splice-sites that mediate the excision of the alternative *Hoxa9* intron are conserved across all Vertebrates (data not shown). U1: binding site for the U1 snRNP, which forms part of the eukaryotic spliceosome, recognizing 5'-splice sites.

Leukemia (AML), despite its inability to bind DNA (Stadler et al. 2014). These authors engineered a version of the *Hoxa9* gene that does not undergo alternative splicing, and show that this *Hoxa9* version, forced to produce a version of the gene that includes the Homeodomain, significantly reduced the leukaemogenicity of this *locus* (Stadler et al. 2014). As such, it becomes clear that the control of the differential RNA processing of *Hoxa9* is important, as the production of a *Hoxa9* isoform that does not encode for a Homeodomain is evolutionary conserved, developmentally regulated and sufficient for the leukaemogenicity of *Hoxa9*.

I thus inquired into the mechanism of *Hoxa9-Homeodomain-less* production by differential RNA processing. To study this, I first compared the reference and *Homeodomain-less* mRNA isoforms of *Hoxa9*, in both mice and humans. I find that in both animal models, the evidence shows that *Hoxa9-Homeodomain-less* mRNAs differ from the reference *HD-containing* isoform due to the exclusion of a 173-ribonucleotide region encoding for a portion of the N-terminal domain (**Figure 4.5**). This exclusion is mediated by an alternative splicing that acts on consensus intronic splice sites (a donor GU dinucleotide, as well as an acceptor splice site AG, **Figure 4.5**). This splicing event leads, in turn, to the introduction of a translational frame-shift, as the number of ribonucleotides that is excluded from the *Hoxa9-Homeodomain-less* protein-coding sequence is not a multiple of 3. This frame-shift is responsible for the introduction of a novel codon downstream of the spliced intron (AGT, encoding for a Serine amino acid), which is immediately followed by an early STOP codon. An additional splicing event that occurs downstream, joining the presumptive N-terminal exon with what a 3'-exon that would otherwise encode for the Homeodomain, appears to be constitutive as it is observed in both isoforms. Due to the aforementioned frame-shift however, this constitutive splicing event joins two portions of the 3'UTR in the case of the *Hoxa9-*

Homeodomain-less isoform, and the Homeodomain-encoding region is thus included in the 3'UTR of *Homeodomain-less* forms (**Figure 4.5**).

To better understand this splicing event, I first looked for the conservation of *Hoxa9-Homeodomain-less* alternative splicing sites across animals. Using the UCSC genome browser, I find that both the 5'-donor and the 3'-acceptor splice-sites are ultraconserved, being present in all 100 vertebrates included in the precomputed UCSC *Multiz* alignments (data not shown). This includes the zebrafish (*Danio rerio)*, as well as the lamprey, and shows that the minimal set of *cis*-regulatory sequences that mediate the *Homeodomain-less* splicing are conserved across all vertebrate genomes. I next used the human Splicing Finder tool (Desmet et al. 2009) to scan the human *Hoxa9* locus for additional splicing motifs, such as exonic splicing enhancers (ESEs) and/or silencers (ESSs), as well as a putative branch point - a ribonucleotide, usually adenine, that is responsible for interacting with the donor splice site to create a lariat-like splicing intermediate. This is a crucial step in a successful splicing reaction, as it precedes both the cleavage of the intron at the 3'-splice site, as well as the ligation of exons. Using this approach, I find that the intronic sequence, absent in *Hoxa9-Homeodomain-less*, includes a strong branch point (data not shown). Additionally, many splicing enhancers accumulate just upstream of this intron, being among the strongest predicted splicing enhancers in the whole region. I obtain an identical result when I scan the cDNA for the full *Hoxa9* mRNA isoform (lacking, as such, the second, constitutive intron) using HSF (Desmet et al. 2009). These results suggest that the mRNA sequence for the full *Hoxa9* mRNA contains all the *cis*-regulatory regions that are necessary for the formation of the *Hoxa9-Homeodomain-less* isoform, and led us to the hypothesis that *Hoxa9-Homeodomain-less* could be formed simply by the excision of the 173 bp intron at the level of splicing.

4.2.5 – The *Hoxa9* mRNA sequence is sufficient for the production of alternative mRNAs that lack the Homeodomain.

To test the hypothesis that *Hoxa9-Homeodomain-less* mRNAs are formed by the excision of an intronic region in *Hoxa9-HDcontaining* mRNAs, I first started a HEK293-EBNA cell culture in the host Laboratory (generously provided by Souvik Naskar in Guy Richardson's Laboratory at Sussex University). At 70% confluence in T75 flasks, HEK293 cells were split and seeded in 6-well plates, and left to rest overnight. The following day, I prepared plasmid-lipid complexes at room temperature in a sterilized fume hood, mixing Lipofectamine 3000 and 2 μg of plasmid DNA in Opti-MEM (reduced serum medium) at a ratio of 3:1 (m/v). The plasmid DNA used was a commercially available vector containing an untagged human *Hoxa9-full* cDNA clone, downstream from a Cytomegalovirus (CMV) promoter – hereafter referred to as *pCMV-Hoxa9*, see Chapter 2 and (**Figure 4.6A**). To assess the efficiency of transient transfection, I performed an identical protocol using the same amount of a GFP-containing plasmid under the control of a CMV promoter (*pCMV-GFP*) and observed the results on an inverted fluorescent microscope. I found that 70-90% of cells expressed GFP at both 12h and 24h after transfection (**Figure 4.6B**).

I thus transfected HEK293 cells with the *pCMV-Hoxa9* plasmid using the aforementioned protocol, and assessed the cellular expression of this vector by first performing timed RNA isolations at 0h, 3h, and 16h after transfection, in three independent biological replicates. These samples were then subjected to cDNA production and PCR, using two sets of primers: one primer pair for the amplification of a region in the GAPDH mRNA (used as a reference gene), as well as a second primer pair that span the excised alternative *Hoxa9* intron in the *Homeodomain-less* mRNA

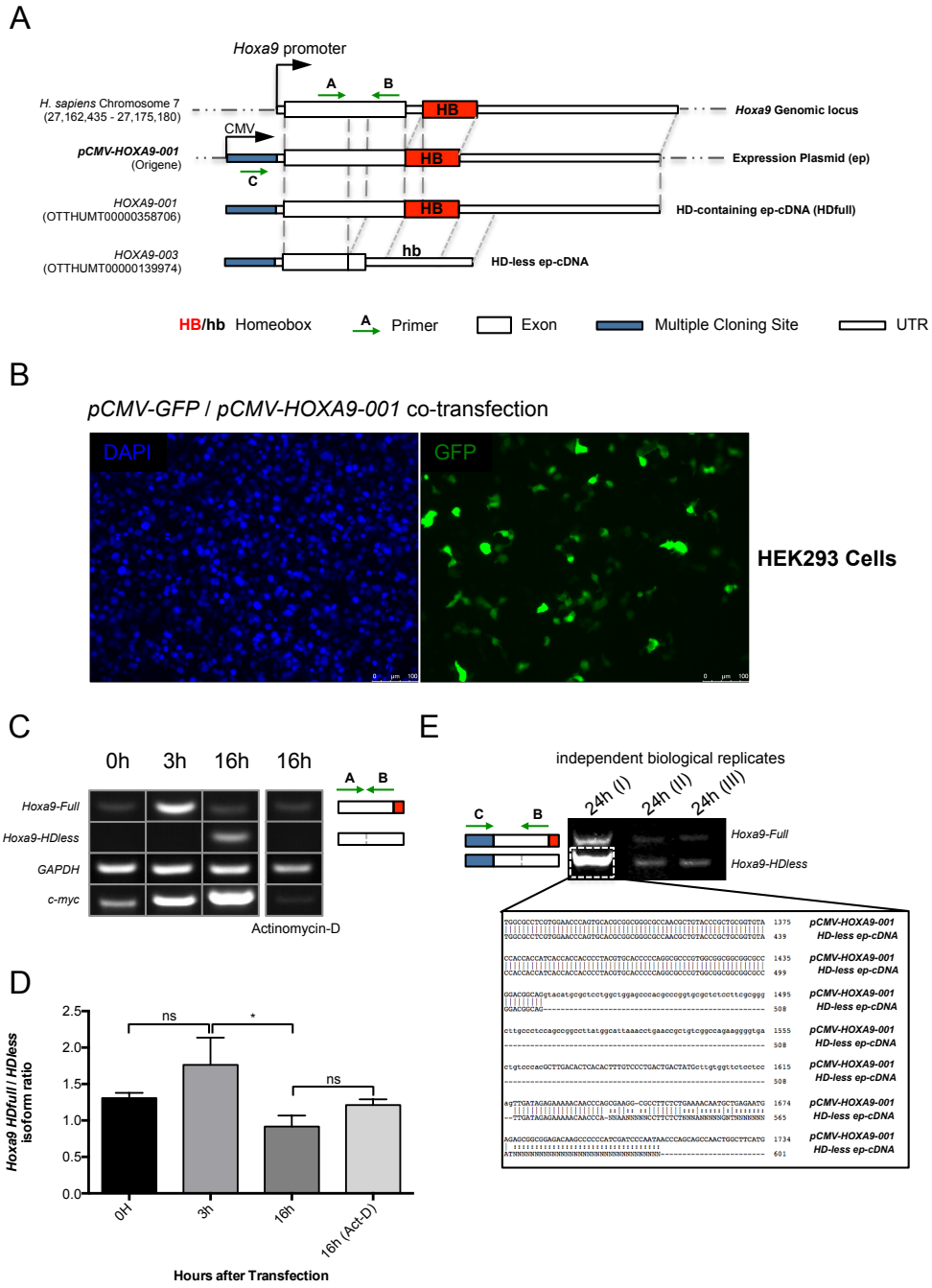


Figure 4.6 – The Homeodomain-encoding cDNA of *Hoxa9* is sufficient to produce the Homeodomain-less mRNA upon overexpression (legend in the following page)**.**

Figure 4.6 – The Homeodomain-encoding cDNA of *Hoxa9* is sufficient to produce the Homeodomain-less mRNA upon overexpression. (A) *Hoxa9* construct used in HEK293-EBNA transfections. This construct contains the cDNA of the Homeodomain-containing *Hoxa9* isoforms (see **Figure 4.5**), lacking the *Hoxa9* constitutive intron. **(B**) Co-transfection of human HEK293-EBNA cells with the *pCMV-GFP* and *pCMV-Hoxa9-001* plasmids. Expression of GFP was observed upon transfection of a CMV promoter-driven GFP construct, along with a *pCMV-Hoxa9* plasmid (in equal molar ratios). Most transfections had an efficiency of 70-90%. **(C**) Timed expression of *Hoxa9* after transfection of HEK293-EBNA cells with the *pCMV-Hoxa9-001* plasmid. Following transfection of the *pCMV-Hoxa9-001* in HEK293-EBNA cells, I observe an initial accumulation of the Homeodomain-encoding *Hoxa9* isoform after 3 hours, which then decreases in expression at 16 hours, as the *Hoxa9* Homeodomain-less isoform accumulates. Upon treatment with Actinomycin-D at 3 hours, I see that this dynamic expression of alternative *Hoxa9* isoforms is abolished, pointing to a link between the differential RNA processing of *Hoxa9* and transcriptional input. Note that untransfected cells express small amounts of the Homeodomain-encoding *Hoxa9* isoform, but not the Homeodomain-less mRNA. **(D**) Quantification of differential RNA processing of *Hoxa9* after transfection. The observed qualitative switch in the outcome of *Hoxa9* differential RNA processing (between 3 and 16 hours) is statistically significant in three independent biological replicates (p= 0.0093, ratio paired t-test) **(E**) Differential RNA processing of mRNAs from the *pCMV-Hoxa9-001* plasmid. To confirm that the *pCMV-Hoxa9-001* plasmid produces the Homeodomain-less *Hoxa9* isoform, I performed an expression analysis of our RNA samples using a forward primer that is complementary to the multiple cloning site, and a reverse primer that anneals downstream of the alternative *Hoxa9* intron. In three independent replicates, I observe that the *pCMV-Hoxa9-001* plasmid produces the Homeodomain-less isoform(confirmed by sequencing).

isoform. The latter primer pair is expected to recognize both the *Hoxa9-full* and the *Hoxa9-Homeodomain-less* isoforms, but the corresponding amplicons vary in size, with the *Hoxa9-Homeodomain-less amplicons* being 173 b.p. shorter that the *Hoxa9-full*-derived amplicons (**Figure 4.6C**).

I find that upon transient transfection with a *pCMV-Hoxa9* plasmid, products of the expected size are detected for *Hoxa9-full* in three biological replicates. Although untransfected cells (0h) also show expression of *Hoxa9-full*, the expression of this isoform is much higher at 3h in transfected cells, showing a decrease in expression at 16h. In the case of the *Hoxa9-Homeodomain-less*, no expression is detected in untransfected cells. In transfected cells however, I see a small but clear expression of *Hoxa9-Homeodomain-less* in HEK293 cells. At 16 hours after transfection, the expression of *Hoxa9-Homeodomain-less* is maximal, becoming the dominant *Hoxa9* isoform. These results show that the expression of a *Hoxa9* mRNA isoform that does not encode for a Homeodomain is detected in HEK293 cells and suggests that the *cis*-regulatory sequences that are present in the cDNA of the *Hoxa9-full* isoform are sufficient to produce the *Hoxa9-Homeodomain-less* within 3 hours post-transfection (**Figure 4.6C-D**).

The fact that untransfected HEK293 cells express *Hoxa9* mRNAs (*full* variant) introduced the possibility that the *Homeodomain-less Hoxa9* isoform could also have been produced endogenously after transfection, possibly by transcriptional activation of the endogenous *Hoxa9* locus by either the plasmid Hoxa9 protein or one of its targets. To confirm that the *Hoxa9-Homeodomain-less* mRNA isoform was indeed being produced by the *pCMV-Hoxa9* plasmid, I performed PCRs on cDNA samples from *pCMV-Hoxa9* transfected HEK293-EBNA cells, using the aforementioned endogenous *Hoxa9* reverse primer, as well as a forward primer that anneals to the plasmid's multiple

cloning site (MCS) sequence. This sequence is present immediately downstream of the CMV promoter and immediately upstream of the *Hoxa9-full* cDNA (see **Figure 4.6A**), and is as such expected to be transcribed and present in plasmid-derived mRNAs. With this experimental set-up, I confirmed the existence of a plasmid-derived *Hoxa9-Homeodomain-less* band after by sequencing (**Figure 4.6E**). Briefly, two plasmid-derived PCR bands of a size compatible with *Hoxa9-Homeodomain-less* were extracted, purified, premixed with either the MCS-specific forward primer used in PCR amplifications or the reverse endogenous *Hoxa9* primer. These PCR results were then sent to sequencing (performed by Eurofins, please see Chapter 2 for further details). Both reactions yielded sequences that are identical in all aspects to the annotated human *Hoxa9-Homeodomain-less* mRNA sequence, with an exception made for a stretch of sequence upstream of *Hoxa9* that corresponds to the MCS. This result confirmed that the *Hoxa9-Homeodomain-less* mRNA isoform is indeed produced from the *pCMV-Hoxa9* plasmid, rather than resulting from the trans activation of the endogenous *Hoxa9 locus*.

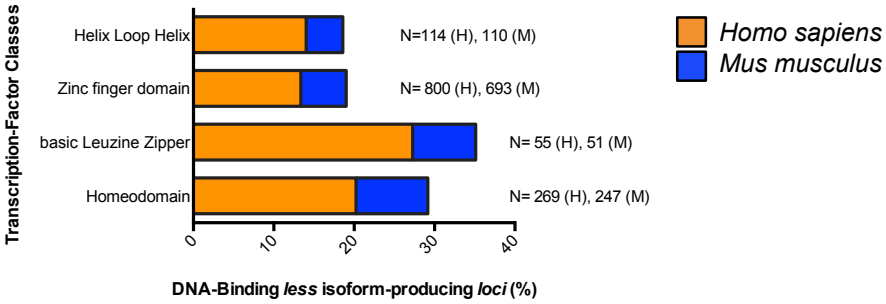
4.2.6 - *Hoxa9* produces mRNAs that lack the Homeodomain in a transcriptional-dependent manner.

Next, I wondered whether the production of a *Hoxa9-Homeodomain-less* mRNA isoform required transcriptional input. I advanced this hypothesis due to three observations, the first by Hatton and colleagues (Hatton et al. 1998), who showed that the *Drosop*hila melanogaster *Hox* gene *Ultrabithorax* (*Ubx)* generates alternative splicing isoforms by a mechanism named *recursive splicing* (or re-splicing), in which a

long intron-containing *Ubx* mRNA isoform is first produced by transcription, being then progressively spliced by intron excision and exon-exon ligation. The second observation comes from our data, as I see that the cDNA sequence of a full-length *Hoxa9* mRNA is sufficient to produce the *Hoxa9-Homeodomain-less* in human cells. Finally, I have observed that the full length, Homeodomain-encoding *Hoxa9* isoform is first produced, as expected, after transfection, but then decreases in quantity as time elapses post-transfection (**Figure 4.6C-D**). Concurrently, the expression of the *Hoxa9-Homeodomain-less* accumulates, as the reference isoform declines (**Figure 4.6C-D**). Together, these observations open up the possibility that an alternative splicing switch might mediate the conversion of full-length *Hoxa9* isoforms into *Homeodomain-less* forms. This is an exciting possibility, as the conversion of one mRNA into another by simple intron excision would provide a rapid switch in gene expression, as all the already-transcribed full-length *Hoxa9* isoforms present in a cell could still be post-transcriptionally converted into *Hoxa9* isoforms that do not encode for the Homeodomain. If this is indeed the case with *Hoxa9*, the production of *Hoxa9-Homeodomain-less* relies on the existence of *Hoxa9-full* isoforms. Furthermore, I expect that if transcriptional activity is blocked 3h after transfection, a time at which I have observed the accumulation of the *Hoxa9-full* isoform and see residual or no expression of the alternative *Hoxa9-Homeodomain-less* isoform, I should nevertheless observe the accumulation of *Hoxa9-Homeodomain-less* isoforms at 16h post-transfection.

However, I observe that in three biological replicates, and upon transcriptional blocking by Actinomycin-D at the 3-hour mark (see Chapter 2), the expression of *Hoxa9-Homeodomain-less* is almost non-existent at 16h. At the same time-point, I observe that a highly unstable control mRNA, *c-myc*, is also seen to decrease in

A



B

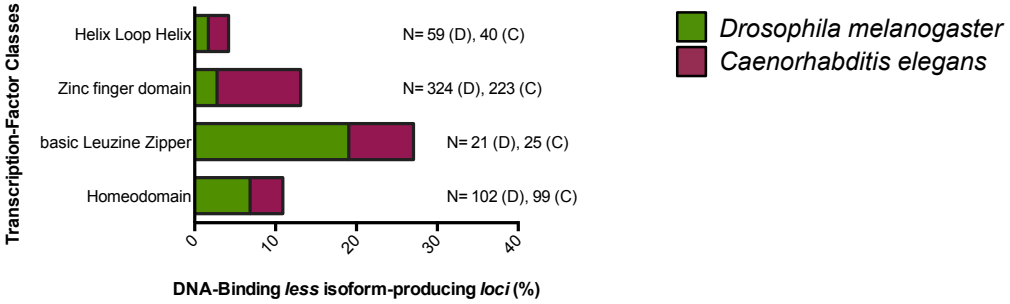


Figure 4.7 – Transcription factors of different classes produce mRNAs that do not encode for the DNA-binding domains across *Metazoa* (legend in the following page)**.**

Figure 4.7 – Transcription factors of different classes produce mRNAs that do not encode for the DNA-binding domains across *Metazoa*. (A) Proportion of loci that produce isoforms lacking the DNA-binding domain across major transcription factor classes in mammals. I used a computational approach to scan the protein sequences of all major mammalian transcription factor classes, looking for evidence of alternative isoforms that lacked the DNA-binding domain (see Text). All major transcription factor classes produce alternative isoforms that lack the DNA binding domain. 20-30% of Homeodomain-encoding loci produce alternative isoforms that lack the Homeodomain in mammals. In the case of basic leucine zipper loci, this number is higher, being close to 30% in *H. sapiens*, and above that figure in *M. musculus*. **(B**) Proportion of loci that produce isoforms lacking the DNA-binding domain across major transcription factor classes in invertebrates. In both *Drosophila melanogaster* and *Caenorhabditis elegans*, I observe that the proportion of loci that produce DBD-less isoforms is less common than in mammals. However, all transcription factor classes have alternative isoforms that lack the DBD in both organisms. These results indicate that differential RNA processing remodels the protein sequences of all major transcription factor classes, and suggests that this regulatory mode could have a significant impact on the transcriptome of bilateral animals.

expression after Actinomycin-D treatment (**Figure 4.6C-D**). Nonetheless, this is not the case with the reference gene *GAPDH*, which is clearly expressed at 16h after Actinomycin-D treatment. This result indicates that the lack of *Hoxa9Homeodomain-less* gene expression is selective, and not just due to apoptosis as a consequence of generalized transcriptional blocking. At 16h post-transfection after transcriptional blocking, I also observe that and the ratio of *full/Homeodomain-less Hoxa9* isoforms is similar to untransfected cells. Together, these results suggest that re-splicing is not the mechanism by which *Hoxa9Homeodomain-less* is formed. Being that as it is, this experiment effectively blocks all transcriptional input in HEK293 cells, and could interfere with the production, by transcription, of endogenous trans-regulators that can mediate the conversion of one *Hoxa9* mRNA isoform into another. Although unlikely, this is certainly a possibility to be experimentally explored in the future.

Re-splicing has been hypothesized as a mechanism by which large introns are progressively removed from pre-mRNAs. Indeed, the *Drosophila melanogaster Ubx* transcriptional unit spans introns that are two orders of magnitude greater than the 173 bp alternative intron of *Hoxa9*. In this case, Hatton and colleagues have speculated that the production of a long *Ubx* mRNA, followed by the progressive excision of a 74 Kb intron in a series of small fragments, effectively reduces the competition among the various *Ubx* splice-sites. If this is the case, recursive splicing would be the exclusive property of genes like *Ubx*, whose transcriptional unit spans approximately 80 Kb. However, a genome-wide study in *Saccharomyces cerevisiae* by Tardiff and colleagues (Tardiff et al. 2006) has shown that at least 90% of yeast genes undergo post-transcriptional splicing. As the average gene length is 1.6 kb in the budding yeast (based on the *Ensembl Fungi* database), this implies that recursive splicing could, in principle, occur in smaller eukaryotic genes like those of mammals.

Together, the results in this section show that the production of the human *Hoxa9-Homeodomain-less* mRNA isoform is independent of *Hoxa9* promoter activity. Additionally, I see that the *Hoxa-9Homeodomain-less* mRNA isoform accumulates only after the full length *Hoxa9* isoform has been produced, and that this accumulation is paralleled by a decline in *Hoxa9-full* expression. This indicates that all the *cis*-regulatory sequences that are required for the alternative splicing of *Hoxa9* are present in the *Hoxa9-full* cDNA. Finally, I observe that a hypothesized post-transcriptional splicing switch, that would convert full length *Hoxa9* mRNAs into *Hoxa9-Homeodomain-less* isoforms is unlikely but not disproven by our experiments.

Our results, as well as the results of others, collectively support a mechanistic model in which a quick, concentration-dependent regulatory switch leads to the production of *Hoxa9-Homeodomain-less* isoforms in the *Hoxa9* locus. This regulatory switch relies on transcription, but not on the specific transcriptional input of the native *Hoxa9* promoter; it also relies on the successful splicing of a cryptic intron, possibly by de-repression of exonic splice-sites. Together, these two levels of differential RNA processing lead to the production of Hoxa9 proteins that do not carry the homeodomain but include amino acids that mediate the interaction of Hoxa9 and co-factors. Additionally, this switch was observed to lie at the interface between proliferation and differentiation cell states, causing the former in the case of human AML (Stadler et al. 2014). I hypothesize that the quick production *Homeodomain-less* proteins plays a role in the regulation of Hox partnership in molecular interactions with co-factors (hetero- and homo-dimers), which can have an indirect effect on their recognition of DNA targets. Thus, this differential RNA processing can have a knock on effect on global Hox target recognition and function, leading to consequences at the level of cellular behaviour in development and disease.

4.2.7 – All major Transcription Factor families produce mRNA isoforms that do not encode for a DNA-binding domain in a conserved manner across metazoans.

I have previously shown that a host of *Hox* genes are expected to produce Homeodomain-lacking protein isoforms in both mice and humans (see section 1 of this chapter). I have also experimentally shown that for at least 5 *Hox* genes, the expression of mRNAs that do not encode for a Homeodomain is observed and appears to be regulated in time and space throughout the development and adulthood of *Mus musculus* (see section 2 of this chapter). Additionally, I have studied the differential RNA processing of the human *Hoxa9* gene, and observed that the production of the *Hoxa9-Homeodomain-less* isoform relies, first, on the initial expression of the *Hoxa9-full* isoform, and then surges in expression as the *Hoxa9-full* isoform becomes less and less preponderant. Together, these results indicate that differential RNA processing generates a number *Hox* isoforms that lack the Homeodomain, and that this differential RNA processing activity appears to be regulated.

To understand whether this is a *Hox*-specific RNA processing routine, I next asked whether I could observe evidence for the production of *Homeodomain-less* isoforms in other Homeodomain-containing genes that lie outside the *Hox* gene family. To this end, I retrieved the identifiers for all GENCODE-annotated Homeodomain genes for both mice and humans, using the Homeodomain SMART ID as a filter (SM00389). This yielded a list of 247 genes in humans, and 269 genes in mice. I then used this list of genes to retrieve all GENCODE-annotated alternative protein isoforms for these genes, as well as the InterProScan domain predictions attached to each alternative Hox protein. Using this approach, I find evidence of widespread production

of alternative isoforms that do not encode for the Homeodomain in both humans and mice. Briefly, I see that 50 genes produce a total of 77 *Homeodomain-less* isoforms in humans, with 24 Homeodomain genes producing 30 *Homeodomain-less* isoforms in mice. Interestingly, 12 of these *loci* are the orthologous between mice and humans, indicating that, and in the previously reported cases of *Hoxa1*, *Hoxa9*, *Hoxb1*, *Hoxb9* and *Hoxc4*, the production of *Homeodomain-less* isoforms is conserved across the mammalian lineage.

To expand on these results, I performed a similar analysis in both *Drosophila melanogaster* and *Caenorhabditis elegans*. In these organisms, I respectively retrieved a list of 102 and 99 Homeodomain genes. Interestingly, I find that at least 7 genes produce 9 *Homeodomain-less* isoforms, one of which is *Antennapedia* (*Antp*), a homologue of mammalian *Hox* groups *6-8*. In the case of *C. elegans*, I find that 4 Homeodomain genes produce 6 *Homeodomain-less* isoforms. Given that both the *Mus musculus* gene *Hoxa7* and its *Drosophila melanogaster* homologue, *Antp* produce *Homeodomain-less* isoforms, I wondered whether there were more instances of conservation in the production of *Homeodomain-less* isoforms across metazoans. To this end, I retrieved homology annotations from both *Ensembl* and *HomeoDB* (http://homeodb.zoo.ox.ac.uk), as well as reports from previously published research papers. I find that *bicoid*, a distantly related homologue of *Hoxb3*, produces a *Homeodomain-less* isoform as does its *Hoxb3* homologue in humans. Furthermore, the *Drosophila melanogaster ladybird late* (*lbl*) produces a *Homeodomain-less* mRNA isoform, as does its human homologue LBX2. In the case of *C. elegans*, the Homeodomain-containing dve-1 locus produces a *Homeodomain-less* isoform as does its human counterpart SATB1. Similarly, the *C. elegans* Homeodomain gene ceh-44 locus produces a *Homeodomain-less* isoform, as do the human and mouse counterparts,

respectively CUX1 and Cux1. Between *Drosophila melagonaster* and *C. elegans*, I see that in at least two cases, homologous Homeodomain-containing genes produce *Homeodomain-less* isoforms in at least one case: the *Drosophila homothorax* locus (*hth*), homologous to the *C. elegans* gene unc-62. This result is interesting as *homothorax* is a *Hox* co-factor in *Drosophila*. Another *Hox* co-factor, the human PBX1 gene, is also observed to form a Homeodomain-less isoform.

These results suggest that this mechanism is not restricted to *Hox* genes, being widespread across the Homeodomain-containing *loci* of different species. Second, they point to the extreme conservation of Homeodomain-less producing loci across evolution. Finally, these results show that Hox molecular partners like PBX1 and *hth*/unc-62 also produce Homeodomain-less forms in a number of distantly related species.

Finally, I wondered whether other major Transcription-factor families also produce DNA-binding less isoforms by differential RNA processing. To answer this question, I retrieved all human, mouse, *Drosophila* and *C. elegans* genes with annotated basic leucine zipper, zinc finger and helix-loop-helix DNA-binding domains (SMART IDs SM00338, SM00355 and SM00353, respectively). Next, I retrieved all GENCODE annotated protein isoforms for these genes in humans and mice, as well as Flybase/Wormbase annotated isoforms for *Drosophila* and *C. elegans*, respectively, querying for the presence of a DNA-binding domain in all cases. Interestingly, I note the occurrence of alternative mRNA isoforms that do not encode for the *DNA-binding* domain in all Transcription Factor Families and all species analysed (**Figure 4.7**). However, I find that the cross-metazoan conservation of *DNA-binding less* isoforms is less pronounced when compared to the Homeodomain gene family (**Figure 4.7**). The basic leucine zipper gene family shows 1 locus (*CREM*/*Crem*) with the conserved

production of *DNA-Binding Domain-less* (*DBD-less*) across mammals. In the helix-loop-helix family, three genes (*Ahrr*, *Tfeb* and *Mlxip*) produce *DBD-less* isoforms in both mice and humans. In both Transcription Factor families, no conserved production of *DBD-less* isoforms exists beyond the mammalian lineage. Finally, I see that 7 *loci* of the zinc finger transcription-factor family produce *DBD-less* isoforms in both mice and humans. Of these, one (*ZNF280D/Zfp280d/row*) also produces a *zinc finger-less* isoform in *Drosophila*. The *C. elegans* zinc finger gene sma-9 and its homologous *Mus musculus* zinc finger *locus* Hivep3 also show the conserved production of *DBD-less* isoforms. However, The zinc finger family is 2.2 to 3.2 times more numerous than the Homeodomain family in all four species analysed. As such, I conclude that the production of *Homeodomain-less* isoforms is widespread across evolutionary lineages and extends to genes outside the *Hox* clusters. I also see that the production of proteins that lack a DNA-Binding Domain is expected to extend to other Transcription-Factor Families and is present across metazoans. However, the relative conservation of *Homeodomain-less* isoform production across great evolutionary distances indicates a preponderant role for this differential RNA processing mode in animal development, physiology and evolution. Interestingly, I observe that the Homeodomain gene HAT14 produces an mRNA isoform that does not encode for the Homeodomain in both *Arabidopsis thaliana* and *Solanum tuberosum* (potato). This underlines the pervasiveness of this differential RNA processing outcome across multicellular eukaryotes.

4.3 – Discussion.

Differential RNA processing has been previously shown to lead to the production of Transcription Factor proteins that do not possess DNA-binding domains (DBDs) (Taneri et al. 2004). However, the precise RNA processing events by which these isoforms are generated remained unclear. Previous studies have examined the effects of alternative splicing on the production of mRNA isoforms that do not encode for a DBD (Taneri et al. 2004). Here, I focus on the *Hox* gene family, which encodes for Transcription Factors, which serve key functions the developmental of most animals. The mammalian *Hox* genes *Hoxa1*, *Hoxa9* and *Hoxb6* a have previously been shown to generate mRNAs that do not encode for the homeodomain (Fernandez & Gudas 2009; Shen et al. 1991; Hong et al. 1995; Fujimoto et al. 1998).

In this chapter, I employ an unbiased approach to the study of the effects of *Hox* differential RNA processing on the composition of Hox protein sequences in mammals. Using the MEME motif-finding tool, as well as the hierarchical clustering of MEME results, I observe that this unbiased method successfully recovers key Hox protein domains, like the Homeodomain, the PBC-interacting Hexapeptide and the SSYF transcriptional activation domains. When I inspect the alternative occurrence of these key motifs in alternative Hox isoforms, I find that differential RNA processing introduces significant variation in the Hox proteome. Among this variation, I corroborate previously observed Homeodomain-less protein isoforms for *Hoxa1* and *Hoxa9* but not *Hoxb6*. Additionally, I see that *Homo sapiens* loci *Hoxa10*, *Hoxb1*, *Hoxb3*, *Hoxc11* and *Hoxd12* all produce protein-coding mRNA isoforms that lack the Homeodomain. In *Mus musculus*, I observe that *Hoxa1*, *Hoxa7*, *Hoxa9*, *Hoxb9* and *Hoxc4* produce mRNAs that do not encode for the Homeodomain.

In the case of *Hoxa9*, the production of Homeodomain-less isoforms relies on the excision of an intron that would otherwise form part of the N-terminal region of full-length Hox proteins, an event that also excludes the PBC-interacting Tryptophan amino acid from the final protein sequence. However, this event leaves the MEIS interaction domain (MIM) intact, pointing to the possibility that *Hoxa9* isoforms can bind to MEIS but not PBC proteins or DNA. As such, Homeodomain-less isoforms could act as dominant negatives, by competing with full-length Hox proteins for molecular partners of the MEIS family. I also observe that Homeodomain-less isoforms from *Hox* loci *Hoxa7, Hoxb1, Hoxc4* and *Hoxb9* contain the hexapeptide, indicating that the dominant-negative hypothesis could still hold true for Hox-PBC interactions. In the cases of *Hoxa7*, *Hoxa10*, *Hoxb9* and *Hoxc4*, however, *Hox* loci produce alternative isoforms that lack the Homeodomain but include the SSYF motif, which mediates transcriptional activation (Tour et al. 2005). This indicates that some Homeodomain-less Hox isoforms can promote the initiation of transcriptional activity even in the absence of DNA binding domains, perhaps by binding to co-factors. This is supported by the fact that in at least two cases, *Hoxa7* and *Hoxc4*, Homeodomain-less isoforms contain both the SSYF and the hexapeptide motifs.

I also observe that differential RNA processing introduces variations in paralogous-specific protein sequences from the *Hox10* PG. In these *Hox*, differential RNA processing impact the availability of Hoxa10 N-terminal and M1 domains (Guerreiro et al. 2012), which have been shown to mediate the repression of rib fates in the developing lumbar region of *Mus musculus* (Wellik & Capecchi 2003). In humans, three alternative Hoxa10 isoforms exist: one, the full-length isoform, contains N-terminal, M1, M2 and Homeodomain; a second isoforms contains M1, M2 and the Homeodomain, but lack rib-repressing N-terminal regions (Guerreiro et al. 2012; Chang

et al. 1996); both isoforms are conserved between mice and humans. Finally, a third Hoxa10 isoform includes sequences from the N-terminal domain but lacks the M1, M2 and Homeodomain.

Based on these results, I confirm and expand previous observations that link differential RNA processing to significant changes in the anatomy of *Hox* proteins. Based on the amount and specially the degree in which this level of *Hox* regulation impacts on protein sequences, I suggest that this regulatory level of *Hox* expression should be taken into account in further studies into *Hox* function.

In this chapter, I also explore the differential RNA processing mechanisms that underlie the production of *Hoxa9* isoforms, which do not encode the Homeodomain. Using a cell-culture system, I observe that the cDNA of the longest *Hoxa9* isoform, which encodes for the Homeodomain, contains all the *cis*-regulatory regions that are needed for the production of the alternative, Homedodomain-less isoform. Furthermore, I show that transcriptional input is needed for the differential production of the *Hoxa9* Homeodomain-less mRNA, and that this isoform quickly becomes the dominant *Hoxa9* isoform 16h post-transfection. These results suggest that the production of Homeodomain-less *Hox* isoforms can occur by a quick switch at the co-transcriptional level, in which the *Hoxa9* locus rapidly changes the identity of its dominant mRNA output. As the production of a Homeodomain-less *Hoxa9* isoform has a causal link to Acute Myeloid Leukemia (Stadler et al. 2014), the regulation of this process in healthy hematopoietic bone marrow cells is key to the homeostasis of the tissue. Our work advances the notion that the mis-regulation of differential RNA processing could lead to similar disease phenotypes in at least five additional *Hox* loci. Further work should deepen the understanding of the molecular mechanisms that underlie the production of Homeodomain-less *Hox* isoforms, including which *trans*-regulators affect this pattern,

and which regulatory signals can underlie the switch in differential mRNA production. *Hoxa1* has been shown to produce Homeodomain-less isoforms upon Retinoic Acid activation (LaRosa & Gudas 1988). It would be interesting to see if the full-length *Hoxa1* cDNA is also able to produce a Homeodomain-less isoform, as the differential processing modes by which Homeodomain-lacking isoforms are produced from the *Hoxa9* and *Hoxa1* loci are the same, and conserved across mammals. By deepening our understanding of how this switch in RNA processing works, I expect to expand our understanding of the links between differential RNA processing and human disease.

Chapter V

The role of Hox *3'UTRs in the coordination of spatial gene expression during mammalian development*

5.1 – Chapter Overview

In the previous chapter I showed that alternative routes of RNA processing in *Hox* and other Homeodomain Transcription Factor genes lead to the formation of mRNA isoforms that do not encode for a Homeodomain in a wide range of organisms. Additionally, I took steps into understanding the mechanism by which Homeodomain-less RNA isoforms are formed, using the *Homo sapiens Hoxa9* in a cell-culture system to show that the formation of the Homeodomain-bearing isoform precedes the formation of the Homeodomain-less form, and that this process is likely not due to recursive splicing.

In this chapter, I use a computational approach to inquire whether the concerted RNA-based regulation of a broad set of *Hox* genes has implications for the spatial expression patterns of these genes, and indirectly for their function. Specifically, I computationally address the impact of 3'UTR-mediated regulation on the gene expression patterns of mammalian *Hox m*RNAs. I employ a commonly used computational cladistic method (Subtree Pruning and Regrafting, SPR) to a novel question: can I computationally test whether 3'UTRs contain information that impacts mRNA spatial expression patterns. I perform this by treating both 3'UTR motifs and the corresponding mRNA's spatial expression patterns as characters, and asking if I see any correspondence between the two. I propose that *Hox* 3'UTRs contain information that impacts on host gene expression during the embryonic development of mammalian tissues. I start by focusing on the development of the *Mus musculus* forelimb and show that in this context, *Hox* genes with similar 3'UTR sequences have significantly similar gene expression patterns. I then demonstrate that common ancestry does not explain the match between 3'UTRs and gene expression. I show that these conclusions also extend

to the match between the 3'UTR sequences and expression of *Hox* and other evolutionarily unrelated genes in the mammalian hindbrain. Finally I successfully validate this novel computational approach, using previously reported experimental results on the control of spatial gene expression by 3'UTRs in the germ line of *Caenorhabditis elegans*. Based on these results, I suggest that within developing mammalian tissues, as with the *C. elegans* germline, co-expressed genes share a network of 3'UTR motifs that reflect their expression patterns. I hypothesise that the exposure of distinct mRNAs to similar regulatory microenvironments during mammalian development can lead to convergent evolution of their 3'UTR sequences.

5.2 – Results

5.2.1 – Mammalian *Hox* 3'UTR contain a host of shared, conserved sequence motifs.

The developing forelimb of *Mus musculus* presents an excellent opportunity for the study of a relationship between 3'UTRs and gene expression patterns. In this tissue, the expression of all 20 *Hox* genes from two genomic clusters (A and D, see chapter I) is spatiotemporally complex and has been shown to affect both the growth and the structural patterning of the limb (reviewed in (Zakany & Duboule 2007)). Additionally the remaining 19 mammalian *Hox* genes belonging to cluster B and C appear to have residual or no expression at all in this tissue, and the deletion of each of these clusters leads to no significant phenotype in adult limbs (reviewed in (Zakany & Duboule 2007)). The 3'UTRs of this group provide, as such, a good opportunity for a negative control in the search for forelimb-relevant motifs.

I first performed an unbiased search for shared sequence motifs in the 3'UTRs of the 20 *HoxA/D* genes with well-known forelimb-bud expression patterns, using the

motif-discovery tool MEME (*positive sequences* set, see chapter 2). I elected this method over regular 3'UTR alignments for four different reasons: first, I often required a measure of sequence similarity between non-homologous sequences - or very distantly related, in the case of the *Hox* 3'UTR analysis in the forelimb – and in this case, a simple sequence alignment, which assumes homology, is inadequate; second, 3'UTRs accumulate, on average, more mutations than coding-sequence regions, being notoriously hard to align without syntenic considerations, and even when considering two orthologous 3'UTR sequences gene within the same genus (Patraquim et al. 2011); third, as with transcription factor-binding sites, two stretches of sequence with identical *cis*-regulatory information are not required to have the respective motifs in the same 5'-3' order in *cis*, as mRNA secondary structure, microRNA-binding and RBP-binding motifs can display a modular and thus relatively position-independent nature; finally, I used motif-finding rather than alignments as motifs such as the aforementioned examples need not be perfectly conserved to have the same regulatory readout, being often degenerate at the base-pair level, while conserving identity at the level of RNA secondary structure (Macdonald 1990) or affinity to a *trans*-regulator.

To enrich our analyses in 3'UTR motifs that are relevant for the forelimb expression context, I performed a discriminative analysis using the 3'UTRs of the 19 *HoxB/C* genes that display little impact on forelimb development as a *negative sequences* set. This means that the analysis would only return 3'UTR motifs that are simultaneously present in one or more 3'UTRs of the 20 *HoxA/D* genes whose expression impacts on forelimb development, and absent in the 3'UTRs of the remaining 19 *Hox* genes. Along with each set of mouse sequences, I submitted the 3'UTRs of the respective *Homo sapiens* homologues; this conservation criterion was expected to increase the sensitivity of our analysis to truly functional motifs. Together

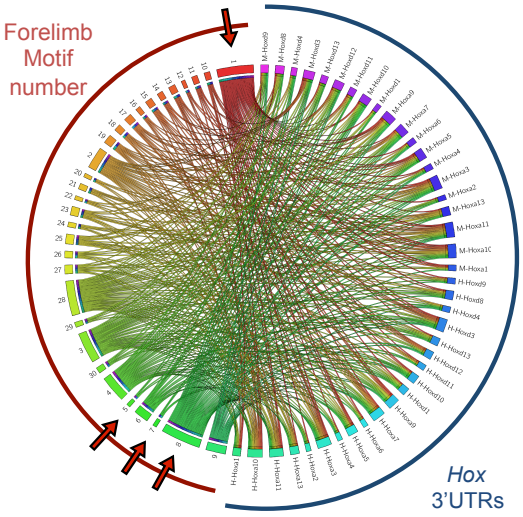
with the submission of the closely related *HoxB/C* sequences as a *negative* set, the inclusion of human sequences is expected to decrease the signal-to-noise ratio, and thus enrich our results in forelimb-relevant motifs. I find a host of shared, conserved and degenerate motifs of 37 nucleotides in length on average, which are shared between the 3'UTRs of forelimb-expressed *Hox* genes in mammals. As a control, the query sequences were shuffled and subjected to the same analysis, yielding no significant MEME motifs. A summary of the results is shown in Figure (**Figure 5.1**).

Next, I wondered whether there was any positional bias in the location of motifs along the *Hox* 3'UTRs. To address this, I first normalized all 3'UTR lengths to the longest 3'UTR in the dataset (human *Hoxa13*, 3180 b.p. long), and then recorded the 3'UTR location of each motif for the 40 sequences included. I find that forelimb-enriched motifs are preferentially located in the initial 40% portion of *Hox* 3'UTRs, and depleted in their distal portion (around the 80% mark into the normalized 3'UTRs) **(Figure 5.1C)**. When compared with a dataset that includes all 3'UTR motifs (obtained by submitting the same sequences through an identical MEME analysis, this time with no *negative sequences* set), I find that while the enrichment in the beginning of the 3'UTRs is present, the distal depletion in motifs appears to be lost. I performed a negative control for motif position-bias by submitting the shuffled 3'UTRs of *Hox* genes through the same analysis. In this case, both proximal enrichment and distal depletion disappear (**Figure 5.1C**). These results indicate that 1) *Hox* genes tend to accumulate *cis*-motifs next to the beginning of the 3'UTR and 2) that forelimb-enriched 3'UTR sequences specially depleted in distal portions of *Hox* 3'UTRs, suggesting that alternative polyadenylation is not expected to radically remodel the set of *Hox* 3'UTR *cis*-regulatory sequences that are deployed in forelimb cells.

Next, I hierarchically clustered both the 20 *Mus musculus Hox* 3'UTRs and

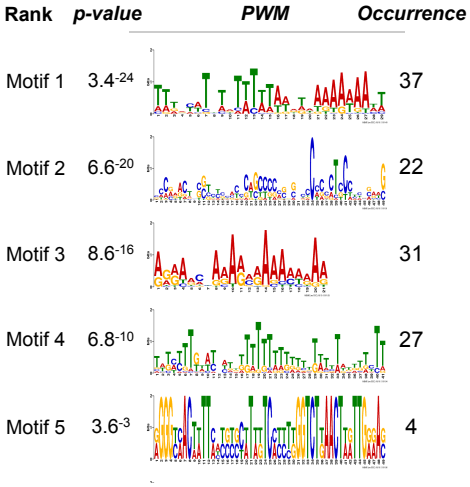
A

HoxA/D genes share a host of conserved 3'UTR motifs



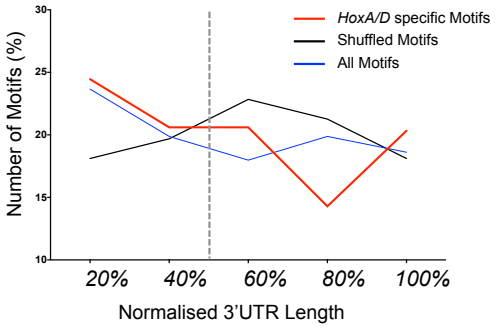
B

Top 5 conserved *Hox* 3'UTR Motifs



C

Proximal regions of *HoxA/D* 3'UTRs are enriched in conserved motifs



D

Size distribution of Forelimb *Hox* 3'UTR Motifs

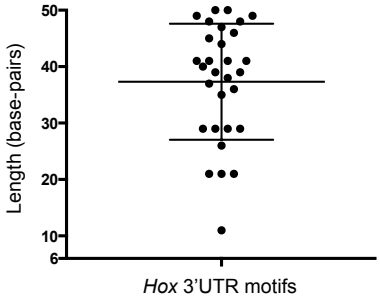


Figure 5.1 – *HoxA/D* genes share a large number of conserved 3'UTR motifs (legend in the following page).

Figure 5.1 – *HoxA/D* genes share a large number of conserved 3'UTR motifs. (A) Circular diagram showing the shared network of *cis*-regulatory motifs among the 3'UTRs of mammalian *HoxA/D* genes, following a discriminative MEME search. I used the 3'UTRs of *HoxB/C* genes of both *H. sapiens* and *M. musculus* as a negative sequence set. As such, the 30 3'UTR motifs that resulted from this analysis are enriched in *HoxA/D* 3'UTRs, and do not occur in *HoxB/C* 3'UTRs. The longest annotated *Hox* 3'UTR was used in all analyses. Motif numbers denote their rank in the query results. Red arrows point to Motifs 1, 5, 7 and 8. Note that Motifs 1 and 8 are shared by most *HoxA/D* genes, while motifs 5 and 7 are shared by few *Hox* 3'UTRs, despite ranking highly in the analysis **(B**) Position weight matrix of the five top-ranking motifs in the discriminative *HoxA/D* 3'UTR analysis. The four highest-ranking motifs are shared by 22-37 of the 40 *Hox* genes analysed, while motif 5 is present in only four 3'UTRs. PWM: Position Weight Matrix. **(C**) Proximal regions of *Hox* 3'UTRs are enriched in conserved motifs. The occurrence of all 30 motifs along the normalized length of *HoxA/D* 3'UTRs (*HoxA/D* specific motifs) is more pronounced in proximal 3'UTR regions, and depleted in distal regions. A non-discriminative MEME analysis for *HoxA/D* 3'UTRs ("All motifs"), recovers the observed proximal 3'UTR enrichment but not the distal depletion of *HoxA/D* 3'UTR motifs. A negative control using a shuffled *HoXA/D* sequence set does not recapitulate any of these results. **(D**) Size distribution of *HoxA/D*-specific 3'UTR motifs. Most 3'UTR motifs are long, averaging 37 nucleotides in size. This result suggests that each motif might not have a one-to-one relationship with a *trans*-regulator *i.e.* these sequences might represent regulatory complexes, as both miRNA and RBP target sites are usually smaller than 10 nucleotides.

the corresponding orthologous human sequences (see Chapter 2), using the presence or absence of motifs as the organizing principle (**Figure 5.2**). I find that with one exception (*Hoxd12*), orthologous 3'UTRs are clustered together, indicating that the initial intention of submitting human and mouse 3'UTRs together was successful in that most of the recovered motifs are conserved. Most of the 30 motifs are present in only 2-5 of the twenty *Hox* genes analysed (**Figure 5.2**); as such, the organization of these genes by 3'UTR motif-similarity offers enough resolution to differentiate between genes with similar 3'UTRs. Additionally, paralogous genes are not found to cluster together (with the exception of *Hoxa1* and *Hoxd1*), and as such do not share the 3'UTR motif-complements uncovered by our analysis. This result suggests that many of these 3'UTR motifs may have arisen after gene duplication. The fact that *Hox* clusters A and D are hypothesized to be evolutionarily closer to clusters B and C, respectively, than each other, supports this idea as the original *HoxA/D* cluster divergence would have happened at the base of the vertebrate lineage (Soshnikova et al. 2013). This would provide enough evolutionary time for the 3'UTRs of paralogous genes to diverge, as the 3'UTRs of non-paralogous but co-expressed genes converge. I hypothesize that this pattern is the outcome of shared expression patterns, leading to shared selection pressures that occur at the molecular level within the different regulatory microenvironments of complex developing tissues.

5.2.2 – Shared *Hox* 3'UTR motifs significantly match mRNA co-expression profiles in the mouse forelimb.

To test whether Limb-expressed *Hox* genes with similar 3'UTR motifs also share spatial expression patterns, I first retrieved previously published models representing the

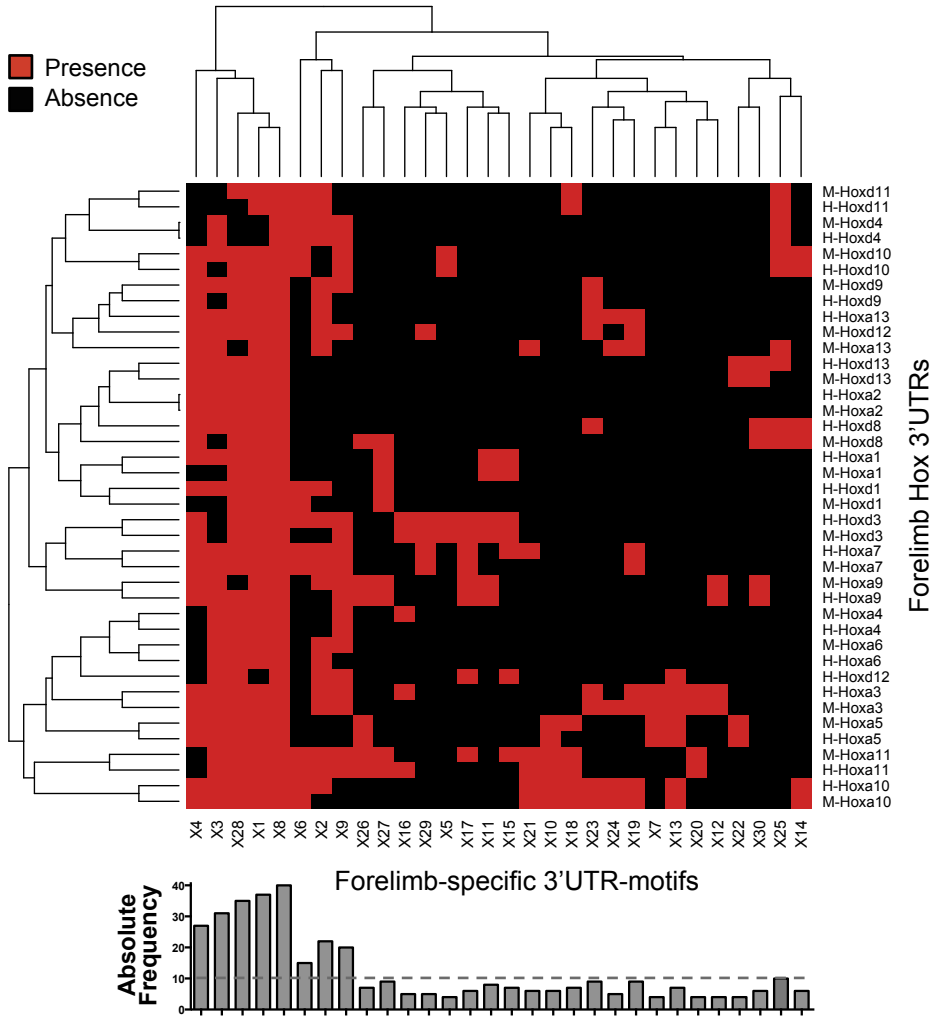


Figure 5.2 – Hierarchichal clustering of *HoxA/D* genes based on shared 3'UTR motifs (legend in the following page)**.**

Figure 5.2 – Hierarchichal clustering of *HoxA/D* genes based on shared 3'UTR motifs. Hierarchical clustering analysis of mammalian *HoxA/D* genes based on shared 3'UTR motifs. I used the results of a MEME discriminative query of *HoxA/D* 3'UTR motifs (see **Figure 5.1**) as informative characters to cluster *HoxA/D* genes based on their similarity in 3'UTR sequences. I find that, with the exception of *Hoxd12*, all *HoxA/D* orthologues of *M. musculus* and *H. sapiens* are clustered together. This result confirms that our query includes ultra-conserved 3'UTR sequence elements. Note that, with the exception of PGs 1 and 13, most paralogues do not occur together, suggesting that our motif dataset offers enough resolution to distinguish between the 3'UTRs of closely related genes. The bottom panel shows the absolute frequency of each motif across *HoxA/D* 3'UTRs. Note that most motifs are shared by the 3'UTRs of a few *Hox* genes, while some MEME motifs are present in the 3'UTRs of most *HoxA/D* mRNAs.

spatial mRNA distribution of all 20 mouse *Hox* genes in the developing forelimb (Zakany & Duboule 2007), (**Figure 5.3)**. I decided to use these models as available images of RNA *in-situs* for all 20 genes were either lacking or of very heterogeneous quality, and could thus compromise the analysis (see Chapter 2 for a fuller explanation). Two stages of *Hox* gene expression were analysed, deemed *early* and *late* (Zakany & Duboule 2007). The forelimb expression patterns of all 20 *HoxA/D* genes are shown in (**Figure 5.3B)** (adapted from (Zakany & Duboule 2007)). Comparing the expression patterns of different genes, it becomes clear that many of these genes have a complex temporal and spatial expression profile, with aspects of this profile being shared between evolutionarily distant genes. As an example, *Hoxa1* and *Hoxd13* share a small anterodistal domain of co-expression in the early stages of forelimb bud development (see **Fig 5.3B)**; the most recent common ancestor of the *Hoxa1* and *Hoxd13* loci lies deep in the metazoan phylogeny, before the emergence of the *Bilateria*).

I transformed each gene's spatiotemporal expression profile into a binary profile (a sequence of 672 data points with either "1" or "0" numerical values), and used these to hierarchically cluster the 20 *Mus musculus HoxA*/D genes based on their shared expression patterns during forelimb development (**Figure 5.3C**).

The organization of *Hox* genes based on either their expression or their 3'UTRs can be represented by the vertical phylograms (hereafter deemed *trees*, for simplicity) seen to the left of the heat maps in figures **5.2** and **5.3C**. In order to compare the two trees, I extracted the topology of each, and manually inputted both topologies into a TNT software script suggested and adapted for our purposes by our collaborator Martín Ramirez (**Figure 5.4A**). TNT (Tree analysis using New Technology) is a cladistic software used for phylogenetic analysis (Goloboff et al. 2008). Importantly, this software implements the Subtree Pruning and Regrafting (SPR) tree

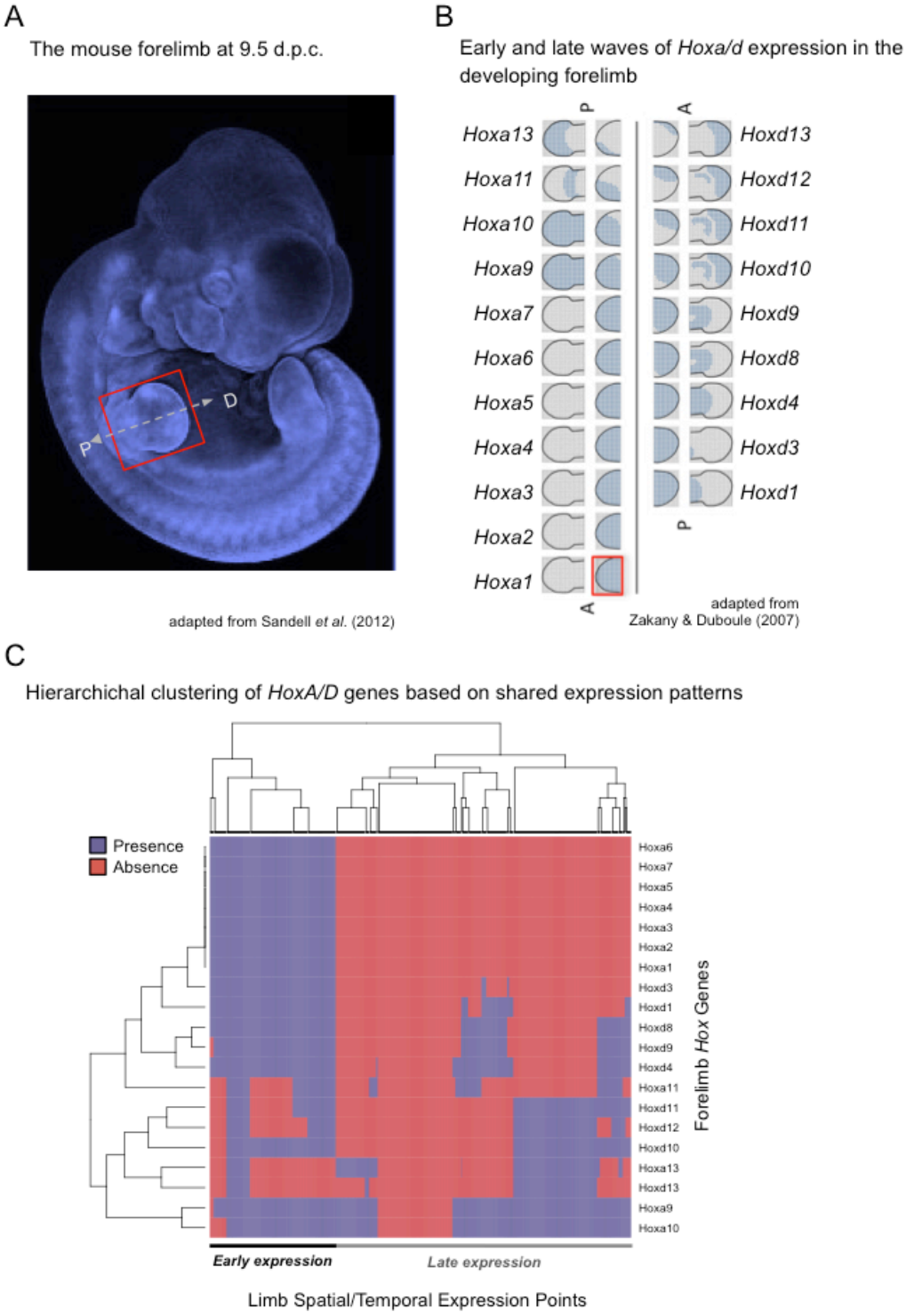


Figure 5.3 – Hierarchichal clustering of *HoxA/D* genes based on co-expression patterns in the developing forelimb bud (legend in the following page)**.**

Figure 5.3 – Hierarchichal clustering of *HoxA/D* genes based on co-expression patterns in the developing forelimb bud. (A) 9.5 d.p.c. *Mus musculus* embryo (adapted from (Sandell et al. 2012)). The red rectangle highlights the embryonic forelimb bud. “P” and “D” indicate the proximal-distal axis. **(B**) Expression diagrams of early and late phases of *HoxA/D* expression during *Mus musculus* forelimb-bud development, showing the spatial distribution of expression for individual *HoxA/D* genes (adapted from (Zakany & Duboule 2007)). Note that genes *Hoxa1-7* are expressed in the whole limb bud in the early expression phase, and show no expression in later stages. This is mostly true for one paralogue of this group of genes (*Hoxd3*), but most paralogues have divergent expression patterns. These expression models were used to build individual gene expression profiles, consisting of a sequence of 672 data points with either “1” or “0” numerical values, corresponding to the presence or absence of *Hox* expression in a specific space and time within the developing limb bud. **(C**) Hierarchical clustering of *HoxA/D* forelimb bud expression profiles. This analysis recapitulates important aspects of the expression models detailed in (B), *e.g. Hoxa1-7* expression profiles are clustered together with *Hoxd1*.

rearrangement method, used to heuristically search for optimal tree topologies in phylogenetic analyses with a large degree of uncertainty (see Chapter 2). In our case, the treatment of both 3'UTR motifs and spatial expression patterns in space as characters (with two possible values, "1" and "0") makes it possible to use this method for the purpose of understanding whether 3'UTRs and expression patterns are significantly similar. This is also possible as both trees have the same clades i.e. the genes are the same, the question being how similarly *arranged* they are.

I found that the *HoxA/D* 3'UTR motif tree significantly matched the expression tree of the same genes in the developing forelimb ($p<0.001$), (**Figure 5.4B**). Furthermore, the 3'UTR motif *tree* had to "pruned" and "rearranged" by SPR (hereafter referred to as *moved*, for the sake of simplicity) 11 times in order to be identical to the gene expression tree (**Figure 5.4B**). As a negative control, I created an array of 10000 random trees using the 3'UTR tree as a starting-point, and compared each with the expression tree. On average, random 3'UTR trees needed 16.6 moves to arrive at the expression tree (**Figure 5.4B**). None of the random trees was as successful as the original tree in matching the expression patterns (**Figure 5.4B**). This result shows the original 3'UTR dataset is exceptionally good at capturing the expression pattern commonalities of *Hox* genes in the forelimb, and demonstrates that organizing *Hox* genes by their forelimb-enriched 3'UTR motif similarities or by expression pattern similarities leads to statistically indistinguishable results .

To test whether these shared motifs were independently acquired in *Hox* 3'UTRs during the evolution of these sequences, and are not just a passive consequence of common ancestry, I performed a phylogeny of *HoxA/D* using protein sequences (see Chapter 2) and used SPR to test whether the shared history of *HoxA/D* genes, typified by their protein phylogeny, could explain their shared forelimb-enriched 3'UTR motifs

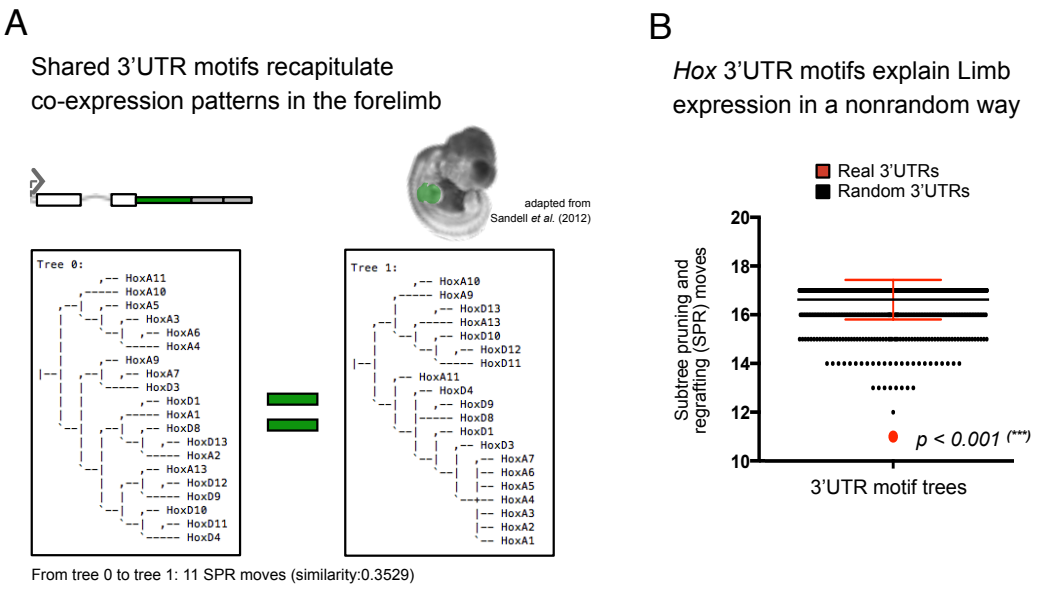


Figure 5.4 – Shared *HoxA/D* 3'UTR motifs significantly recapitulate dynamic *HoxA/D* co-expression patterns in the forelimb (legend in the following page)**.**

Figure 5.4 – Shared *HoxA/D* 3'UTR motifs significantly recapitulate dynamic *HoxA/D* co-expression patterns in the forelimb. (A) Shared 3'UTR motifs match *HoxA/D* co-expression patterns in the developing forelimb. The relative organization of *HoxA/D* genes based on shared 3'UTR motifs (see **Figure 5.2**) was compared with the relative organization of the same genes based on co-expression profiles in the developing forelimb (see **Figure 5.3**) using the SPR method of phylogenetic tree comparison (see Chapter 2 and text). The conversion of the 3'UTR-tree into the expression-tree requires 11 SPR "moves". **(B)** The organization of *HoxA/D* by shared 3'UTR motifs is statistically identical to the organization of the same genes by co-expression patterns. A null distribution of 3'UTR-based *HoxA/D* trees, resulting in 10000 random 3'UTR trees was compared to the co-expression tree in (A).,None of the random trees is as successful as the real 3'UTR-based tree at matching the co-expression patterns of *HoxA/D* genes (as measured by the number of SPR "moves"). **(C**) To control for phylogenetic effects, I performed a protein phylogeny of the mouse *HoxA/D* genes (Neighbour-Joining, JTT substitution model, (Katoh & Standley 2013)). This phylogenetic tree is expected to mirror the evolutionary history of the *HoxA/D* loci. **(D)** The evolutionary history of *HoxA/D* genes does not explain their co-expression patterns in the developing forelimb. The phylogeny of *HoxA/D* genes was compared with the co-expression tree of the same genes (right panel in (A)) and to 10000 randomizations of the *HoxA/D* phylogeny. However, the real phylogenetic relationship of *HoxA/D* clusters was not better than random trees at matching co-expression patterns.

(**Figure 5.4C-D**). I found the answer to statistically unsupported (p=0.228, NS), as in order to reconstruct the 3'UTR-based tree, the phylogenetic tree of *HoxA/D* genes had to be *moved* 16 times, where the average moves from 10000 null trees based on the original *HoxA/D* phylogenetic tree to our 3'UTR-motif tree was 16.62 (**Figure 5.4D**). This result indicates that the independent evolution of different *Hox* 3'UTRs after gene duplication, rather than their shared ancestry, best explains the match between combinatorial *Hox* 3'UTR motif complements and *Hox* expression patterns in the mouse forelimb.

5.2.3 –Forelimb-enriched *Hox* 3'UTR motifs include RNA secondary-structures, as well as RBP binding-sites.

To generate hypotheses about the mechanism by which these 3'UTR motifs are expected to mediate the spatial restriction of *Hox* mRNAs in the forelimb, I first used the MEME suite TOMTOM tool (Bailey et al. 2009) to compare our forelimb *Hox* motif-set with the recently-published RNA-binding motif dataset of Ray and colleagues (Ray et al. 2013). I find that many motifs show statistically significant matches to RBP-binding motifs (**Figure 5.5**). Interestingly, the RBP HuR is the most common hit, being significantly predicted to bind to 7 different motifs, three of which in the top-five hits (Motifs 1, 4, 5, 8, 14, 27 and 30 – the motif numbers indicate significance rank) (**Figure 5.5A**). This protein is ubiquitous but has nevertheless an enriched expression in branchial arches, neural tube and limb buds at 10.5 d.p.c. (Gouble & Morello 2000), a developmental timing that coincides with the outgrowth and axial specification of the forelimb. Additionally, mutants for HuR present clear limb malformations (Katsanou et al. 2009), accompanied by both the decreased stability and polysomal occupancy – a proxy for translational activity - of a *Hox* mRNA (*Hoxd13*) in the 12.5 d.p.c. Forelimb.

Consistently with this result, *HoxD13* has 4 *HuR* binding motifs in our dataset (Motifs 1, 4, 8 and 30), which are conserved between mice and humans, indicating that HuR could mediate the stability of *Hoxd13* in the developing forelimb through an interaction with the 3'UTR.

Based on the overrepresentation of HuR binding-sites in our forelimb enriched *Hox* 3'UTR motif-set, I wondered whether HuR-binding 3'UTR motifs were sufficient to significantly recover the expression patterns of *Hox* genes in the forelimb. To address this, I repeated the hierarchical clustering analysis of *Hox* 3'UTR motifs, only this time exclusively using the predicted HuR-binding motifs (Motifs 1, 4, 5, 8, 14 and 30). Using an SPR analysis, I find that HuR 3'UTR motifs fail to match the spatiotemporal expression patterns of the host genes ($p=0.228$, data not shown), indicating that HuR-binding is not expected as the only 3'UTR-mediated process that influences the spatial distribution of *Hoxd13* in the forelimb.

The 10.5 d.p.c. upsurge in HuR expression in the developing limb buds is paralleled by that of AUF1 (HNRNPD) (Gouble & Morello 2000). This RBP was also shown to bind the same targets as HuR (Barker et al. 2012), but its binding was seen to have the opposite effect of HuR binding on mRNA stability and translation, as in the case of AUF1 both regulatory processes decrease (Barker et al. 2012). It is thus possible that the competition between AUF1 and HuR for *Hox* 3'UTR targets can affect the spatial distribution of *Hox* mRNAs in the developing forelimb. HuR-binding has been shown to be more efficient in single-stranded RNA sequences (Barker et al. 2012), while AUF1 has been hypothesized to bind and remodel local RNA structures (Wu et al. 2013) due to its RNA chaperone-like activity (Zucconi et al. 2010; Wilson et al. 2003). As such, the outcome of a competition between mouse forelimb-expressed HuR and AUF1 for *Hox*

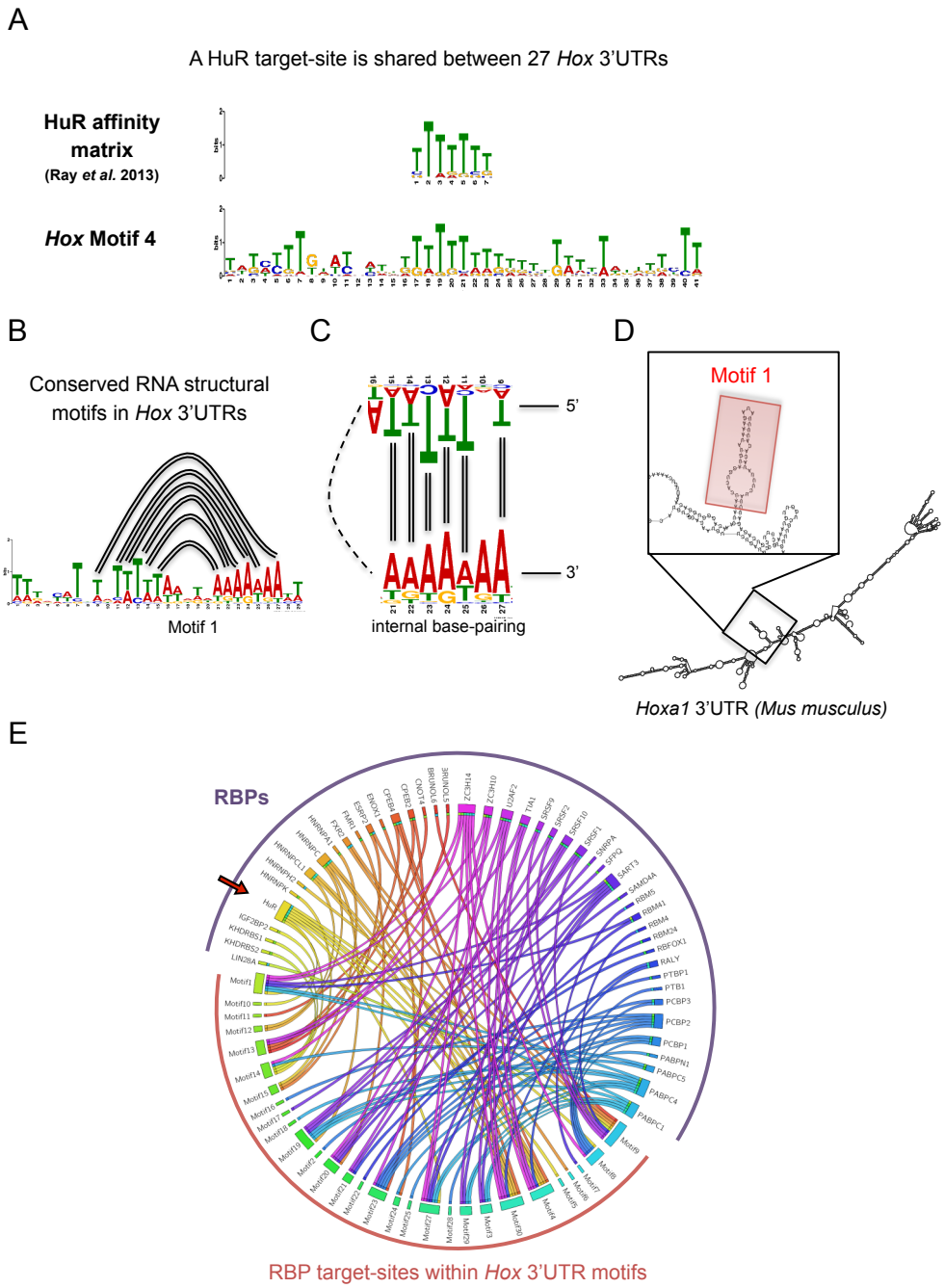


Figure 5.5 – *HoxA/D* 3'UTRs contain numerous RBP-target motifs (legend in the following page)**.**

Figure 5.5 – *HoxA/D* 3'UTRs contain numerous RBP-target motifs. (A) TOMTOM analysis of Motif 4. I used the TOMTOM tool of the MEME suite (Bailey et al. 2009) to look for RBP targets within our set of *HoxA/D* 3'UTR motifs (see **Figure 5.1**) using a recently-published RBP-target dataset (Ray et al. 2013). With this approach, I find that a number of motifs contain predicted RBP-binding sites. Among these is a top-ranking motif for HuR, an RBP with high expression in the developing forelimb, which has been shown to target *Hox* mRNAs (see text). **(B-D**) *HoxA/D* MEME Motif 1 contains a RNA secondary structure motif. **(B)** The most represented 3'UTR motif (Motif 1, see **Figure 5.1B**) contains two complementary stretches of nucleotides separated by a group of degenerate sites. **(C**) This motif can form a secondary RNA structure by internal base-pairing **(D**) RNAFold predictions of RNA secondary structure predict that, in many cases, this 3'UTR motif folds by internal base-pairing in a context-independent manner (see text). **(E)** Catalogue of RBP target sites within *HoxA/D* 3'UTR motifs. Using the TOMTOM tool (see panel (A)), I find that the 3'UTRs of *HoxA/D* genes contain a number of shared RBP targets. Targets for HuR (red arrow) are represented in our dataset, as they occur in motifs other than Motif 4. These results suggest that HuR has the potential to regulate the expression of a cohort of *HoxA/D* genes *via* direct RBP-3'UTR interactions.

3'UTR targets could be determined, in part, by the targets' RNA structure.

Further probing the 30 Motif sequences, I noticed that the top-ranked motif (Motif 1) shows an initial 5-7 base-pair U-rich region, followed by 7 base-pair long A-rich sequence (**Figure 5.5B-D**). These are the most conserved base pairs in Motif 1's 29-nucleotide stretch. This motif is predicted in our analysis to be bound by HuR (p=0.006), and thus by AUF1, and is detected in all except one (*Hoxd4*) of the 20 mouse *Hox* 3'UTRs analysed. To verify if the two U-rich and A-rich stretches could pair to form an internal stem-loop structure in Motif 1, I performed *RNAFold* predictions for the individual Motif 1 sequences of all 20 *Hox* genes in both mice and humans (see Chapter 2). Additionally, I tested for context-insensitivity by performing similar RNA folding predictions in the full 3'UTRs and full mRNAs of the host genes. I find that in the cases of *Hoxa1, Hoxa2, Hoxa3, Hoxa4, Hoxa5, Hoxa11* and *Hoxd1*, Motif 1 is predicted to form a conserved stem-loop, regardless of the sequence context, while in the remaining 13 genes this structure is not observed.

AUF1 has also been shown to promote Argonaute 2 (AGO2)-mediated degradation of *Hoxb8* and other mRNAs in human cells, and is hypothesized to do so by induction of RNA structure changes that expose 3'UTR miRNA target-sites to the AGO2/miRNA complex (Wu et al. 2013). I thus asked whether miRNA targets were present in or near Motif 1, and scanned our 3'UTR motif dataset for target-sites for miRNAs mmu-miR-199a-5p and mmu-miR-214, which are expressed in the developing mouse forelimb (Lee et al. 2009), but found no antiparallel matches to the seeds of either miRNA.

5.2.4 – 3'UTR motifs of evolutionarily unrelated genes match spatial mRNA expression in the mouse hindbrain.

In the previous section, I showed that a combination of 30 motifs that are shared and specific to 20 *HoxA/D* 3'UTRs significantly matches the combinatorial expression patterns of the same genes in the developing forelimbs of mice. Given that these genes are very similar, sharing genomic context, recent common ancestry, sequence features, molecular roles and expression patterns in different tissues, as well as transcriptional and epigenetic regulation, I wondered whether the aforementioned result might reflect a *Hox*-specific regulatory feature. To address this, I applied a similar "3'UTR *Vs*. expression" analysis in a context where a number of very distinct genes are co-expressed, and for which there is experimental evidence of spatially restricted expression patterns: the hindbrain (**Figure 5.6A**).

As the hindbrain of mice first becomes segmented at around 9.5 d.p.c (see Chapter 1), I chose a developmental time window that coincides with rhombomere compartmentalization (8.5-10.5 d.p.c.) to ask whether the segmental restriction of genes to particular rhombomeres (e.g. *Hoxb1*) is reflected in their 3'UTR motif composition. To do so, I first used the MGI-Mouse Gene Expression Database (GXD, see chapter 2) to retrieve a list of 32 *Mus musculus* genes for which there is RNA *in-situ* hybridization evidence of rhombomere-restricted expression. This list of genes includes leucine-zipper, zinc finger and Homeodomain transcription factors, as well as cell-signalling molecules from the BMP, Shh and Wnt pathways (see **Table 5.1**). I also recovered a list of 60 genes whose expression at 8.5-10.5 d.p.c. transgresses rhombomere boundaries.

Second, I used MEME to apply an unbiased and discriminative motif-search in the 32 *Mus musculus* 3'UTR sequences of rhombomere-specific genes, using the 60 3'UTR sequences of genes that are expressed in 2 or more rhombomeres as a *negative sequences* set (see Chapter 2). I found a set of 30 degenerate motifs that are shared between these genes (**Figure 5.6B**). I then hierarchically clustered the 32 genes based

Table 5.1 - A list of 32 genes with rhombomere-restricted expression in the Mus musculus hindbrain.

Associated Gene Name	Interpro ID	Interpro Short Description	Ensembl Family Description
ASCL1	IPR011598	bHLH_dom	ACHAETE SCUTE HOMOLOG 1
BHLHE40	IPR011598	bHLH_dom	CLASS E BASIC HELIX LOOP HELIX CLASS B BASIC HELIX LOOP HELIX ENHANCER OF SPLIT AND HAIRY RELATED SHARP
BMPR1B	IPR003605	TGF_beta_rcpt_GS	BONE MORPHOGENETIC RECEPTOR TYPE
CASZ1	IPR007087	Znf_C2H2	ZINC FINGER CASTOR HOMOLOG 1 CASTOR RELATED
CYP26A1	IPR001128	Cyt_P450	CYTOCHROME P450 EC_1.14.-.-
CYP26C1	IPR001128	Cyt_P450	CYTOCHROME P450 EC_1.14.-.-
DBH	IPR005018	DOMON_domain	PRECURSOR EC_1.14.17.-
DBX1	IPR009057	Homeodomain-like	HOMEOBOX DBX1 DEVELOPING BRAIN HOMEOBOX 1
EBF2	IPR002909	IPT	TRANSCRIPTION FACTOR EARLY B CELL FACTOR
FJX1	IPR009581	DUF1193	FOUR JOINTED BOX 1 PRECURSOR FOUR JOINTED HOMOLOG
GLI1	IPR007087	Znf_C2H2	ZINC FINGER
GSX1	IPR009057	Homeodomain-like	GS HOMEOBOX HOMEOBOX GSH
HOXB1	IPR009057	Homeodomain-like	HOMEOBOX HOX B1
HOXD3	IPR009057	Homeodomain-like	HOMEOBOX HOX
IGDCC3	IPR013151	Immunoglobulin	IMMUNOGLOBULIN SUPERFAMILY DCC SUBCLASS MEMBER PRECURSOR
JUN	IPR002112	Leuzip_Jun	TRANSCRIPTION FACTOR AP 1 PROTO ONCOGENE C JUN
KCTD11	IPR003131	T1-type_BTB	BTB/POZ DOMAIN CONTAINING KCTD11
LHX1	IPR009057	Homeodomain-like	LIM/HOMEOBOX LHX1 LIM HOMEOBOX 1 HOMEOBOX LIM 1
MAB21L1	IPR024810	Mab-21_dom	MAB 21
NHLH1	IPR011598	bHLH_dom	HELIX LOOP HELIX 1 HEN 1 NESCIENT HELIX LOOP HELIX 1 NSCL 1
NKX2-2	IPR009057	Homeodomain-like	HOMEOBOX NKX 2 HOMEOBOX NK 2 HOMOLOG B
PAX2	IPR009057	Homeodomain-like	PAIRED BOX PAX
PAX5	IPR009057	Homeodomain-like	PAIRED BOX PAX
PCP4L1	UNKNOWN	UNKNOWN	UNKNOWN
POU3F2	IPR009057	Homeodomain-like	POU DOMAIN CLASS 3 TRANSCRIPTION FACTOR BRAIN SPECIFIC HOMEOBOX/POU DOMAIN BRAIN
PRTG	IPR003961	Fibronectin_type3	IMMUNOGLOBULIN SUPERFAMILY DCC SUBCLASS MEMBER PRECURSOR
SFRP1	IPR020067	Frizzled_dom	SECRETED FRIZZLED RELATED PRECURSOR SFRP
SHH	IPR000320	Hedgehog_signalling_dom	HEDGEHOG HEDGEHOG N HEDGEHOG C
SP5	IPR007087	Znf_C2H2	TRANSCRIPTION FACTOR
SPRY1	IPR007875	Sprouty	SPROUTY HOMOLOG SPRY
TLX3	IPR009057	Homeodomain-like	T CELL LEUKEMIA HOMEOBOX HOMEOBOX HOX
WNT1	IPR005817	Wnt	WNT

on their 3'UTR motifs (**Figure 5.6C**) or their expression patterns (**Figure 5.6D**), and compared the two trees using SPR (see Chapter 2). As with the case of the comparison between *Hox* 3'UTR combinatorial information and forelimb *Hox* expression similarities, I find that the 32 hindbrain-expressed genes analysed share a set of 3'UTR sequences that, when considered together, is sufficient to match the co-expression patterns of rhombomere-restricted genes ($p<0.0001$) (**Figure 5.6C-D**). Furthermore, the 15 SPR *moves* necessary to remodel the 3'UTR tree in order to reconstruct the expression tree were never observed in any of the 10000 random 3'UTR trees generated and compared with the expression tree; the average number of moves necessary for a random 3'UTR tree to match the real expression tree was 30.9, almost the double amount of remodeling required for the real 3'UTR dataset.

In summary, in the hindbrain as with the forelimb case, organizing a set of genes by 3'UTR motif similarities or by expression pattern similarities leads to statistically indistinguishable results. These results indicate that the 3'UTR motif composition of hindbrain-expressed genes either influences of is influenced by the concomitant expression patterns of the same genes in the hindbrain. Furthermore, it shows that 3'UTR motif sharing in relation to a developmental context is not a property of evolutionarily related genes, and that our computational technique can be applied to more than one developmental context.

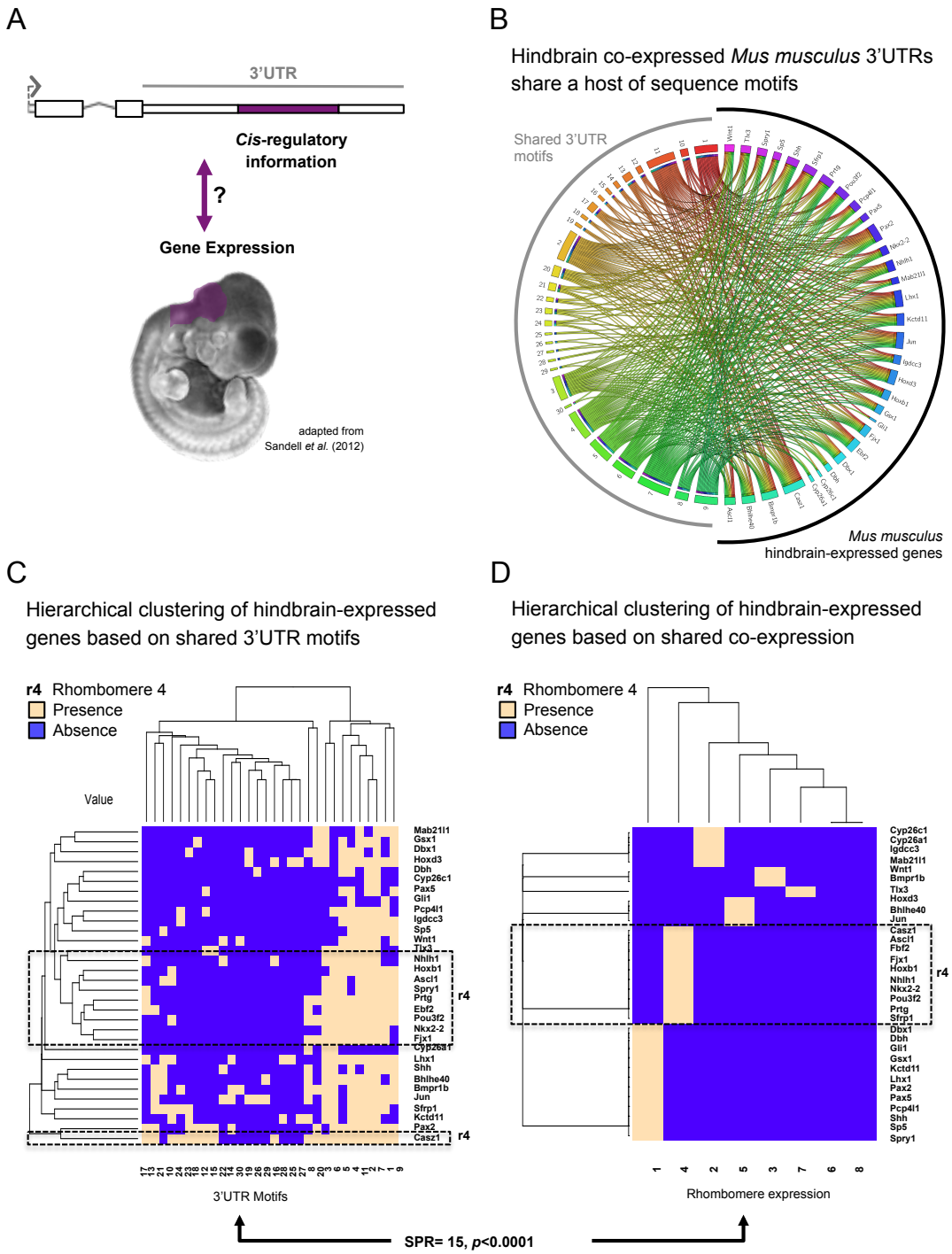


Figure 5.6 – The 3'UTRs of phylogenetically unrelated genes share *cis*-motifs that significantly recapitulate their expression patterns in the *Mus musculus* hindbrain

(legend in the following page).

Figure 5.6 – The 3'UTRs of phylogenetically unrelated genes share *cis*-motifs that significantly recapitulate their expression patterns in the *Mus musculus* hindbrain. (A) Comparison of 3'UTR motif complements and co-expression patterns of hindbrain-expressed genes. **(B)** Circular diagram showing the shared network of *cis*-regulatory motifs among the 3'UTRs of hindbrain-expressed genes after a discriminative MEME search for shared 3'UTR motifs in *Mus musculus* genes with rhombomere-restricted expression patterns (see text and **Figure 5.1)**. The 3'UTRs of hindbrain-expressed genes that transgress rhombomere boundaries were used as a negative sequence set. Despite being evolutionarily unrelated, hindbrain co-expressed genes contain a host of shared 3'UTR motifs. **(C)** Hierarchical clustering of hindbrain-expressed genes based on shared 3'UTR motifs. This analysis results in the grouping of a number of genes that are co-expressed in rhombomere 4 (r4, dashed box) – compare with panel D. **(D)** Hierarchical clustering of Hindbrain-expressed genes based on shared expression patterns. Most r4-expressed genes are captured by the analysis based on 3'UTR similarities (dashed box, compare with dashed box in (C)). An SPR-based comparison of both organizations shows that this set of 32 hindbrain-expressed genes ashare a set of 3'UTR sequences that are sufficient to match the co-expression patterns of rhombomere-restricted genes ($p<0.0001$). This result suggests that the 3'UTRs of a number of evolutionarily unrelated genes contain *cis*-regulatory sequences which relate directly to specific regulatory contexts during the development of mammals.

5.2.5 – Validation of the SPR method as a test for 3'UTR-mediated coordination of gene expression: the *Caenorhabditis elegans* germline.

In previous sections, I use the cladistic tree-comparison SPR method to show that clustering genes by their shared 3'UTR motif-complements is statistically the same as organizing them by gene expression patterns. This is true in both the embryonic forelimb and hindbrain of *Mus musculus*. In this section I validate of our methodology using the *C. elegans* dataset in (Merritt et al. 2008).

The authors report a large number of 3'UTRs that, when fused to a reporter construct, successfully limit the spatial expression patterns of the host protein in the germline (Merritt et al. 2008). Briefly, the authors chose 30 genes known to be expressed in the germline of *C. elegans* in a spatially regulated manner, and cloned their 3'UTR sequences downstream of a GFP fused to Histone H2B. The *pie-1* promoter was used to drive the blanket expression of all constructs in all regions of the *C. elegans* germline. The authors show that for 24 of the 30 genes, the 3'UTRs successfully recapitulate the expression patterns of the host genes in this tissue, and conclude that UTRs are the primary regulators of gene expression in the *C. elegans* germline (Merritt et al. 2008). I reasoned that this would be an good set of results with which to validate our computational approach matching 3'UTRs to expression patterns (**Figure 5.7A**).

To this end, I first recovered the reported patterns of expression of the 24 GFP:H2B-3'UTR fusions in the *C. elegans* germline, summarized in the Figure 2 of (Merritt et al. 2008). Second, I retrieved the 3'UTRs of the aforementioned 24 genes, and submitted these sequences to a MEME motif search (see Chapter 2). This generated a set of 30 motifs shared between the 3'UTRs of the 24 germline-expressed genes probed (**Figure 5.7B**). Thirdly, I separately performed two hierarchical clustering

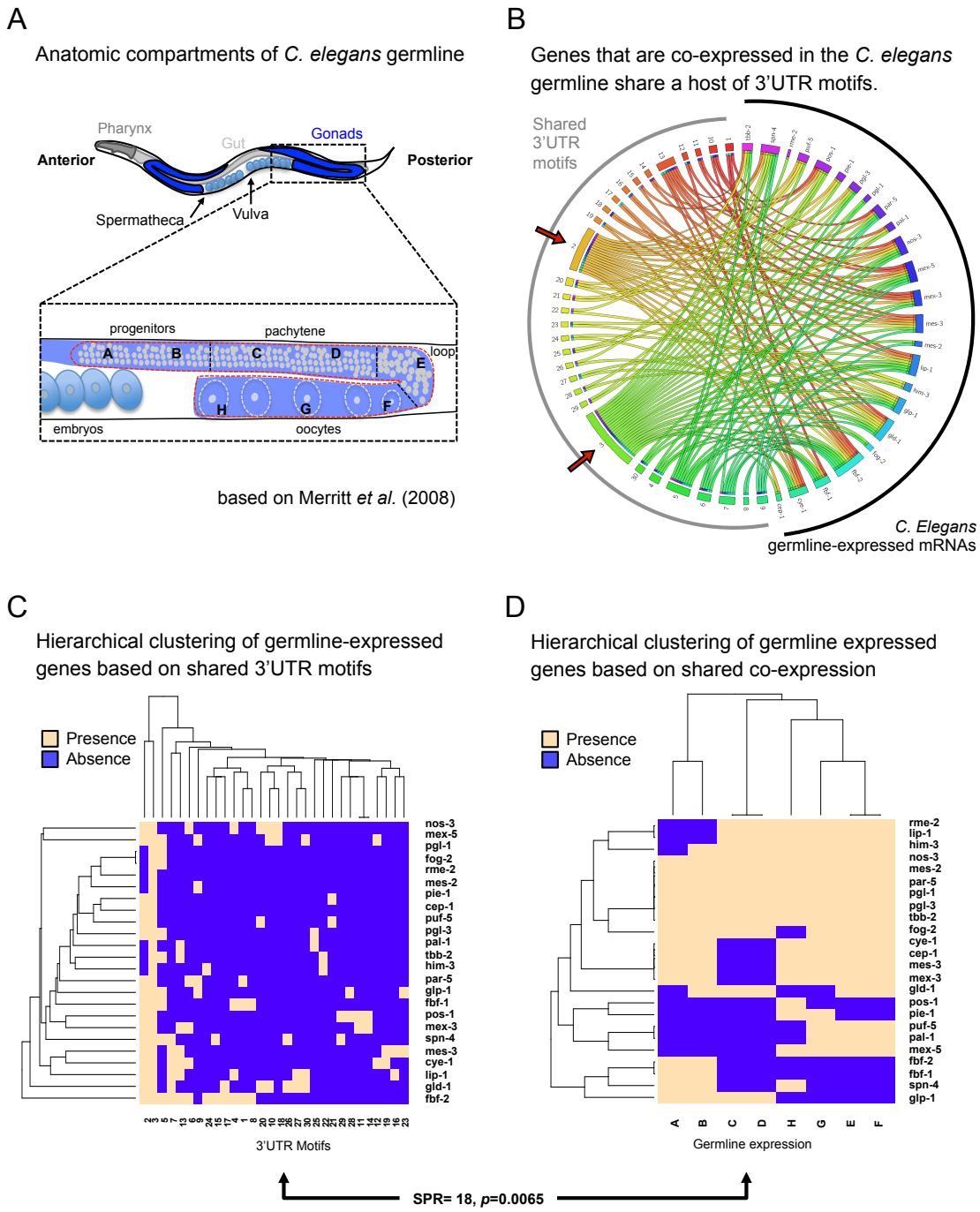


Figure 5.7 – The 3'UTRs of phylogenetically unrelated genes share *cis*-motifs that significantly recapitulate their expression patterns in the *C. elegans* germline (legend in the following page).

Figure 5.7 – The 3'UTRs of phylogenetically unrelated genes share *cis*-motifs that significantly recapitulate their expression patterns in the *C. elegans* germline. (A) Diagram depicting the anatomical organization of a *C. elegans* hermaphrodite, emphasising distinct regions within the germline (A-H). To validate our computational method for the study of 3'UTR *cis*-regulatory information in the context of gene expression, I used the *C. elegans* germline context, where 3'UTRs are the primary regulators of gene expression (Merritt et al. 2008; Reinke 2008). I retrieved the longest annotated 3'UTRs for 24 germline-expressed genes, and performed an unbiased MEME motif search on this sequence-set **(B)** Circular diagram showing the shared network of *cis*-regulatory motifs among the 3'UTRs of germline-expressed genes in *C. elegans*. **(C)** Hierarchical clustering of germline-expressed genes based on shared 3'UTR motifs. **(D)** Hierarchical clustering of germline-expressed genes based on shared expression patterns. The 3'UTR-based tree matches the expression tree significantly better than random trees (p=0.0065).

analyses: one of the 24 *C. elegans* genes in question clustered according to their shared 3'UTR motifs (**Figure 5.7C**), and the second using the same genes, but clustering them according to their shared expression patterns (**Figure 5.7D**). I find that the two trees are significantly similar (p=0.0065), with 18 SPR moves being necessary to match the original trees (**Figure 5.7C-D**). In contrast, 10000 randomized 3'UTR trees needed, on average, 24 moves to perform the same operation.

Together, these results validate our computational approach, and suggest that it can be can be applied to, first, isolate shared developmentally-relevant 3'UTR sequences, second, to ask whether this information is relevant for the regulation of gene expression in a given tissue, and finally, to generate hypotheses about which *trans*-regulators can be responsible for the coordination of this 3'UTR-mediated process.

5.3 - Discussion

In this chapter, I employ a computational approach to explore how the RNA-based regulation of a number of *Hox* genes can influence their transient spatial expression patterns in complex, developing mammalian tissues. I find that in the case of the developing mouse forelimb, *Hox* 3'UTRs share a host of evolutionarily conserved motifs. This network of shared 3'UTR motifs is statistically the same as the co-expression network for the same genes in this tissue. Additionally, I observe that this is not restricted to *Hox* genes or the forelimb, as I observe a similar correspondence between a network of shared 3'UTR motifs and the spatial expression of 32 genes in the developing hindbrain. Unlike the forelimb analysis, these have very different evolutionary histories and molecular functions, but similarly carry 3'UTR sequences that match gene co-expression patterns. Finally, I successfully validate our method using a previously published dataset that reports the 3'UTR-mediated control of gene

expression in the context of the *C. elegans* germline.

The female germline is a classic context for the study of post-transcriptional gene regulation in animals, as transcription is mostly silent in this tissue due to direct inhibition of RNA Pol II (*Drosophila* and *C. elegans*) or chromatin regulation (*Mus musculus*). In *Drosophila melanogaster*, the mRNAs of *bicoid* (Macdonald & Struhl 1988; Macdonald 1990), *nanos*, and *oskar* have been shown to be assymetrically localized in oocytes due to sequences in their 3'UTRs (reviewed in (Johnstone & Lasko 2001)). In particular, *nanos* mRNAs have been shown to have spatially-restricted germline expression in *Drosophila* (Wharton & Struhl 1991), zebrafish (*nanos1*), the sea urchin *Strongylocentrotus purpuratus* (Oulhen et al. 2013) and mice (*Nanos3*) (H. Suzuki et al. 2010). In all cases, this seems to be achieved by a combination of 3'UTR-mediated spatial restriction and translational repression of *nanos* mRNAs.

More generally, 3'UTRs have been shown to spatially restrict the expression of broadly expressed reporter constructs in the germline of *Caenorhabditis elegans*, setting-up spatial gene expression patterns that closely resemble those of the proteins encoded by the endogenous mRNAs of at least 24 genes (Merritt et al. 2008). A similar study in *Drosophila melanogaster* found that the expression of a number of mRNAs is spatially restricted in the *Drosophila* germline and that this process is 3'UTR-mediated and involves translational inhibition (Rangan et al. 2008).

Addressing the question of 3'UTR-based regulation in transcriptionally active cells is technically demanding, as most gene expression analysis techniques recover steady-state mRNA expression levels, and cannot thus distinguish between the relative contributions of transcription, mRNA stability and degradation to those patterns. However, the forced expression of 3'UTR sequences in the context of reporter constructs, and its comparison with the corresponding endogenous protein expression

patterns, offers an experimental setup that can powerfully address the question of how important 3'UTR-based regulation is in transcriptionally active cells.

Although no *en masse* studies exist for tissues other than the germline, individual cases that use reporter-fused 3'UTR sequences support the idea that 3'UTR-mediated restriction of spatial gene expression also occurs in transcriptionally active cells. For instance, Myelin basic protein (MBP) transcripts contain two short conserved sequences that are necessary and sufficient for the successful localization of MBP mRNAs to the myelin compartment of oligodendrocytes (Ainger et al. 1997). In *Drosophila,* the 2.3 kb-long 3'UTR of *Ubx* recapitulates the posteriorly-restricted expression pattern of the endogenous *Ubx* protein in the embryonic CNS, when broadly expressed and fused to an mCherry reporter (Thomsen et al. 2010). In mice, the 3'UTR of *Hoxb4* was shown to be necessary and sufficient for the maintenance of the anterior boundary of somitic *Hoxb4* expression in the paraxial mesoderm after 9 d.p.c. (Brend et al. 2003). These two observations suggest that the study of the 3'UTRs of *Hox* genes in their native expression context can shed some light on the possibility of a general case for the 3'UTR-mediated spatial expression control.

Here, I provide a computational approach to address the question of 3'UTR-mediated regulation of spatial gene expression *en masse*. First, our analyses examine a number of sequences that would be very demanding to study *in vivo*, e.g. 32 genes in the hindbrain of mice - a similar *in vivo* approach would involve the creation of at least 32 individual transgenic mouse lines. Furthermore, I argue that this *en masse* approach would be not only important but vital to address the biological impact of this regulatory process; if, as our analysis indicates, the 3'UTRs of co-expressed genes form a network of shared *cis*-regulatory motifs, and the ratio between commonalities and differences in the 3'UTR modules of co-expressed genes directly correlates with their *degree* of co-

expression, this problem requires the study of many genes in parallel.

The problem of studying steady-state mRNA levels is harder to bypass. It seems to us that our results support the 3'UTR-mediated control of gene expression. However, I assume that these sequences are active at the post-transcriptional level. Although it is hard to envision that all 96 3'UTRs analysed in this chapter exclusively regulate the expression of the host genes at the DNA level, it is quite possible that some do, in addition to or even exclusion of their role in RNA-based regulation. However, I consider the discriminative motif-finding approach previously described to partially address this problem, as the use of related *negative* or background sequences is expected to decrease the incidence of spurious sequences in general, be it transcriptional enhancer sequences or 3'UTR motifs that are not important to the developmental context in question.

I see that the co-option of the SPR method to our developmental biology question yields interesting results in all cases. However, I find that the *C. elegans* analysis is less statistically significant than the ones for the Limb and Hindbrain datasets. I hypothesize that this is due to the *C. elegans* analysis being inherently *less* powerful, for two reasons: first, I provided significantly less information into the *C. elegans* analysis, when compared to the previous two datasets; I used 24 individual sequences, with 379 b.p. of average length. In contrast, I utilized 40 3'UTRs with an average length of 1037 b.p., in the Limb analysis, while in the case of the Hindbrain data, 32 3'UTRs were used, averaging 1397 b.p. in size. As such, the *C. elegans* dataset has 17% of the information used in the Limb analysis, and 20% of that used for the Hindbrain test. Additionally, both the Limb and Hindbrain analyses used *negative sequence* sets (19 and 60 3'UTRs, respectively), while no negative set was used in the *C. elegans* analysis. Together, these two lines of evidence suggest that, while still

statistically significant (p=0.0065), the *C. elegans* analysis would need more data in order to be refined.

One shortcoming of our analysis is that I cannot yet recover successful correspondences between individual 3'UTR motifs and specific areas of gene expression. Explicitly mapping regulatory sequences to individual expression domains in a developing tissue would provide powerful hypotheses about development, its evolution, and its relationship to underlying DNA sequences. With our dataset, this would require an exhaustive combinatorial analysis that is beyond the scope of the present study. However, I am able to generate clear hypotheses regarding the concerted regulation of these genes at the RNA-level, for instance about which *trans*-regulators might act on the 3'UTR motifs of different genes, as with the case of the HuR/ARE1 RBP pair. Based on our results, I submit that in the forelimb bud, an interplay between relative HuR/ARE1 levels, the number of 3'UTR binding motifs and their respective RNA structure is responsible for the either setting up, refining or, minimally, maintaining of *Hox* expression patterns. Additionally, I see an enrichment of forelimb-relevant motifs in the proximal 3'UTRs of *Hox* genes, and a proportional depletion of the same motifs in the distal tracts of *Hox* 3'UTRs. I suggest that alternative polyadenylation is not predicted to significantly change the available *cis*-regulatory complement of *Hox* 3'UTRs in the forelimb. As such, our computational study provides directly testable hypotheses about specific aspects of the RNA-level regulation of gene expression in vivo.

The regulation of *Hox* genes has been extensively studied in the forelimb. In this research field, the two main emerging ideas are that chromatin-based regulation of *HoxA/D* transcription, as well as long-range transcriptional enhancer sequences both mediate the setting up of *Hox* expression in this tissue (see Chapter I). Similarly, a

complex transcriptional cascade has been shown to set up the expression patterns of *Hoxb1*, among other genes, in the developing hindbrain (reviewed in (Alexander et al. 2009)). With this in mind, I do not argue that, as with the case of the *C. elegans* germline, 3'UTRs are the main determinant of gene expression in these mammalian tissues. Rather, I reason that once an mRNA molecule is transcribed, it becomes exposed to a regulatory environment that needs to be minimally tolerated, if not canalized in order for the original mRNA to be successfully translated into a protein, which can go on and regulate transcription in the nucleus. That is to say that even in a scenario where transcription single-handedly sets up the expression patterns of *Hox* genes in the developing forelimb of mice, these mRNAs would still have to biochemically associate with some *trans*-regulators and not others in order to be exported from the nucleus, and localized, stabilized and translated in the cytoplasm; in this scenario, one would still expect the mRNA sequences of *Hox* genes to reflect that process.

In summary, our results point to the general impact of 3'UTRs on mammalian gene expression patterns acting not in isolation, but in *coordination* with other levels of regulation to successfully set up the gene expression patterns of developmentally-relevant genes. I show that, as with tissue-specific transcriptional enhancers, mRNAs display tissue-specific 3'UTR motifs. The fact that these motifs are shared across co-expressed mRNAs points to the co-regulation of said mRNAs by trans-regulators like RNA-binding proteins, acting at the post-transcriptional RNA level. Furthermore, I see that common ancestry does not explain the existence of the aforementioned 3'UTR motifs. As such, I hypothesize that the subjection of different mRNAs to the same specific regulatory contexts within a larger, molecularly complex developing tissue, leads to a common selective pressure on mRNAs of different origin, promoting

evolutionarily convergent solutions to the shared problem of RNA regulation, which are reflected at the nucleotide level, in the sequences of 3'UTRs.

Chapter VI

General Discussion

6.1 – General Discussion

The work presented in this thesis provides novel insights on the mechanisms, developmental consequences and evolution of differential mRNA processing in the *Hox* clusters of mammals, and introduces a novel approach for the study of *Hox* post-transcriptional regulation in mammalian embryos, showing that dynamically co-expressed *Hox* genes tend to share sequence motifs in their 3'UTRs.

Hox genes encode a family of transcription factors that are differentially expressed along the A-P axis of developing animals, and provide positional information to serially homologous segments, eliciting the diversification of axial structure and function (Pearson et al. 2005). Mutations in *Hox* genes lead to major abnormalities in body-plan - homeotic transformations - in which one segment of the animal is transformed in identity and function developing into the likeness of another (Mallo & Alonso 2013; Pearson et al. 2005). This observation further sediments the role of *Hox* genes in the specification of the mammalian body plan throughout development, as well as its evolution.

In *Drosophila melanogaster*, previous studies in the host laboratory have highlighted the impact of differential RNA processing in the expression and function of *Hox* genes within the segments of developing embryos (Thomsen et al. 2010; Rogulja-Ortmann et al. 2014). It has been previously shown that the differential RNA processing of *Antp*, *Ubx*, *abd-A* and *Abd-B* is developmentally regulated (Thomsen et al. 2010), leading to the production of short 3'UTRs in the epidermis at early developmental stages, and a subsequent elongation of 3'UTR sequences in later stages of the central nervous system (Thomsen et al. 2010). Previous studies have also highlighted that

mRNAs from the *Ubx* locus are regulated post-transcriptionally by miRNAs (Bender 2008; Ronshaugen et al. 2005) and RBPs (Rogulja-Ortmann et al. 2014), a regulatory interaction which impacts the quantity, quality and position of *Ubx* proteins in the developing CNS of *Drosophila* embryos (Bender 2008). Mammalian *Hox* genes have also been previously shown to produce alternative mRNA isoforms (See Chapter 1). However, the extent and quality of differential RNA processing in mammalian *Hox* clusters has not been explored in depth in the literature.

In this work, I look to extend previous observations on the differential RNA processing of *Drosophila Hox* genes to mammalian *Hox* clusters. I start by collecting available *Hox* mRNA sequences, asking: (*i*) how do mammalian *Hox* genes produce alternative mRNA isoforms? (*ii*) How did RNA processing evolve in the *Hox* clusters of mammals? (*iii*) Is differential RNA processing predicted to affect the post-transcriptional regulation of *Hox* mRNAs by differential visibility to miRNAs? (*iv*) What is the effect of differential RNA processing on the amino acid sequences of Hox proteins? (v) How do 3'UTR sequences relate to the dynamic expression of *Hox* genes in specific developmental contexts? To address these biological questions, I used a combination of computational methods, as well as a human cell culture system, and show that differential RNA processing is a staple of *Hox* genes and is predicted to significantly expand the Hox protein complement.

6.2 – Paralogous *Hox* genes share patterns of differential RNA processing in mammals.

Unlike *Drosophila melanogaster* and most other animals, the mammalian *Hox*

complement consists of thirty-nine genes, divided into not one but four clusters which reside in different chromosomes (Pearson et al. 2005). The composition of the mammalian *Hox* complement is a result of two early rounds of whole-genome duplication at the base of the vertebrate lineage. This is reflected in commonalities of *Hox* composition across clusters: paralogous *Hox* genes occur in similar relative positions within distinct clusters and tend to share sequence motifs, expression patterns and specific molecular functions; there are thirteen *Hox* paralogue groups, with each being composed of 2-4 *Hox* genes that share a single common ancestor sequence in the *Hox* cluster of the chordate common ancestor. Paralogue groups also show a great degree of functional redundancy. For instance, the genes of the *Hox10* and *Hox11* paralogue group of *Mus musculus* are expressed during the morphogenesis of the axial skeleton (see Chapter 1). When mutated, these genes lead to homeotic transformations in which specific sections of the axial skeleton are transformed into the likeness of another (Wellik & Capecchi 2003). This phenotype, however, only becomes apparent when 5 of the 6 copies of the *Hox10* and *Hox11* genes are mutated in *Mus musculus* (Wellik & Capecchi 2003).

In this thesis, I retrieve freely available alternative *Hox* mRNA sequences and study their occurrence across *Hox* clusters (see Chapter 3). I show that the incidence of alternative mRNAs in *Hox* clusters is lower that the transcriptomic average in both *Homo sapiens* and *Mus musculus*. However, when the total mRNA count of each paralogue group is averaged by the number of paralogues in that group, rendering an average rate of alternative isoform production per *Hox* gene in the context of its duplication group, I see that the average incidence of differential RNA processing is similar between *Hox* genes and the rest of the genome. This indicates that the duplication history of *Hox* genes explains the data better than considering each *Hox*

gene in isolation. Moreover, I see that the average production of mRNA isoforms is heterogeneous across paralogous *Hox* groups, being enriched in PGs 3 and 6-9. It would be interesting to explore the alternative mRNA complement of the *Hox* paralogue group 3 in the context of the hindbrain. This vertebrate structure has conserved *Hox* expression patterns between *Homo sapiens* and *Danio rerio* (Alexander et al. 2009). In this context, *Hox3* expression patterns correspond to the sites of origin of neural crest cells (Kiecker & Lumsden 2005) and *Hoxa3* mutations lead to defects in neural crest formation (Chisaka & Capecchi 1991). Neural crest cells are a vertebrate developmental innovation, and are thought to underlie major evolutionary adaptations in this clade (Holland et al. 1994), as they linked to the formation number of features in the vertebrate head, linked to novel modes of predation. Based on our results, I hypothesize that the control differential RNA processing of *Hox3* genes in the hindbrain could underlie some of these developmental adaptations.

I also show that the heterogeneity in average alternative isoforms per paralogue group is conserved between *Homo sapiens*, *Mus musculus*, and *Danio rerio*. However, the incidence of differential mRNA processing is uncorrelated between paralogous genes of different vertebrate *Hox* clusters. This indicates that in different organisms, different *Hox* genes contribute mRNAs to the conserved paralogous isoform pool.

Additionally, I see that there are two main sequences of differential RNA processing events, which dictate the formation of alternative *Hox* mRNAs. These two modes integrate distinct kinds of transcriptional initiation and alternative splicing with the formation of alternative 3'UTRs. In this context, I also find that *Hox* genes of the same paralogue group share a dominant mode of differential RNA processing.

Together, these results suggest that the production of alternative mRNAs is not independent between paralogous *Hox* genes. Rather, I suggest that *Hox* genes of the

same paralogue group contribute to a pool of alternative mRNAs that is shared by paralogues. In the context of redundancy in sequence, expression and function among mammalian *Hox* paralogues, I hypothesize that differential RNA processing generates distinct mRNAs and proteins from similar *Hox* loci, diversifying the protein complement of otherwise similar genetic loci. As most genes were lost after the two rounds of genomic duplication in the early vertebrates, I suggest that this mechanism could underlie the uncommon conservation of redundant *Hox* loci in mammals. This is supported by the fact that the relative patterns of differential RNA production across paralogue groups are conserved between *Homo sapiens* and *Danio rerio*, which indicates that differential RNA processing patterns were present at the base of the vertebrate lineage.

Other authors have proposed a relationship between gene duplication and alternative mRNA production by alternative splicing, postulating that genes with large amounts of alternative splicing are more often maintained after gene duplication, with a resulting subfunctionalization of paralogues by fixation of alternative isoforms (Kopelman et al. 2005; Su et al. 2006; Talavera et al. 2009). This results in the functional complementarity between paralogues, and a secondary loss of alternative isoform production across duplicates to decrease redundancy. Indeed, the size of gene families that result from duplication events are negatively correlated with alternative splicing rates in *Homo sapiens* and *Mus musculus* (Kopelman et al. 2005).

Our evidence supports a functional-sharing hypothesis, but the extent to which mammalian *Hox* clusters have lost alternative isoforms after the early vertebrate rounds of gene duplication is unclear. *Danio rerio Hox* genes produce less alternative mRNAs than their mammalian counterparts; this lineage also underwent an additional round of gene duplication. However, this dataset might be severely depleted, as is evidenced by

the lack of information on *Hox* 5' and 3' mRNA sequence ends. In this context, it would be interesting to extend the study of *Hox* differential RNA processing to the cephalochordates, a sister group to the vertebrates with a single *Hox* cluster composed of 14 paralogues. The observation of a comparatively enriched production of alternative mRNAs across the *Amphioxus Hox* cluster would support the hypothesis that the mammalian rates of differential *Hox* RNA processing are the result of gene duplication and subsequent paralogue subfunctionalization. Conversely, low rates of alternative mRNA production in *Amphioxus Hox* genes would indicate a vertebrate-specific mechanism for the maintenance and functional complementarity of *Hox* paralogues. Recently, some authors have suggested an alternative model for the evolution of alternative splicing after gene duplication, in which genes accumulate alternative splicing isoforms over evolutionary time under relaxed selection regimes, until selection starts acting against the addition of novel isoforms in older genes (Roux & Robinson-Rechavi 2011). In this context, genes with less alternative isoforms are prone to duplicate more frequently (Roux & Robinson-Rechavi 2011). A more complete view of the *Danio rerio* and *Amphioxus Hox* transcriptomes, which underwent respectively one more and no duplications, could lead the way in discerning between these two competing hypotheses.

6.3 – Differential RNA processing diversifies Hox protein-sequences in mammals.

In the previous section I introduce the notion that paralogous *Hox* genes share a pool of mRNAs in mammals, and that the diversification of *Hox* function at the differential RNA processing level might have led to a decrease in the strength of natural

selection against the genetic load brought by gene redundancy. This scenario might explain why a number of highly redundant genes have been maintained in the genome while most duplicated genes were lost in the vertebrate lineage. However, if a pool of exclusive mRNAs from different paralogues were needed to sustain *Hox* function, I expect to find phenotypic changes upon the mutation of a single *Hox* paralogue, as the pool of available paralogous mRNAs would be diminished. As mentioned in the previous section, this is not true in the case of *Hox10* and *Hox11*. It could be that the functional redundancy of paralogous *Hox* genes is manifested in the morphogenesis of broad inter-segmental structures, while more detailed intra-segmental morphogenesis relies on the regulatory input of different paralogues. Indeed, *Hoxa9* and *Hoxd9* display both broad functional redundancy in the morphogenesis of the axial skeleton and the forelimb, but also show specific funtions at a smaller morphogenetic scale within these tissues (Fromental-Ramain et al. 1996).

In our work, I show that differential RNA processing is predicted to significantly remodel the open-reading frames of *Hox* mRNAs (Chapter 4), supporting the idea that together with alternative transcriptional initiation and alternative polyadenylation, alternative splicing has the potential to diversify *Hox* molecular functions. I observe that *Hox* genes of different paralogue groups produce alternative isoforms that lack key *Hox* motifs like the DNA-binding Homeodomain, the protein-protein interaction hexapeptide and the SSYF transcriptional activation domains in a combinatorial and often conserved manner.

Other authors have shown that alternative splicing introduces variation in the availability of DNA-binding motifs of Transciption factors in both *Mus musculus* and *Homo sapiens* (Taneri et al. 2004). Additionally, these authors have shown that the variation which alternative splicing introduces in transcription factor sequences is larger

than that observed for other loci(Taneri et al. 2004). Alternative isoforms also tend to have tissue specificity (Taneri et al. 2004). In addition to this, alternative splicing has previously been shown to generate alternative *Hox* mRNA isoforms that do not encode for a Homeodomain (see Chapters 1 and 4). In the case of *Hoxa9*, differential RNA processing was shown to produce Homeodomain-less isoforms in bone marrow hematopoietic cells of both *Mus musculus* and *Homo sapiens* (Stadler et al. 2014). In *Mus musculus*, the Homeodomain-less *Hoxa9* isoform has been shown to underlie the leukaemogenic potential of the *Hoxa9* locus (Stadler et al. 2014), an observation that underlies the importance of the regulation of differential *Hox* RNA processing in mammals.

Using a cell-culture system, I observe that the longer of the two *Hoxa9* mRNAs, which encodes for the homeodomain, includes all *cis*-regulatory sequences that are necessary for the production of the homeodomain-lacking version. Although I observe a transcriptional dependence for the production of the *Hoxa9-HD-less* isoform, this link is not unexpected (Kornblihtt et al. 2004). I hypothesize that a quick, quantity-dependent switch between different RNA processing modes might underlie the regulation of this process. Additionally, I observe that *Hoxa1* produces alternative mRNAs that do not include the homeodomain in the same coordinated manner, while other genes like *Hoxc4*, *Hoxb1* and *Hoxb9* seem to employ different regulatory steps to achieve the same end. This points to a degree in the plasticity of differential RNA processing, in which many kinds of differential processing reactions might be employed to generate the same combinatorial end. Additionally, I observe that the production of alternative isoforms that lack the DNA-binding domain is observed in all major transcription factor classes, and shows an exceptional degree of conservation, being observed in some cases between mammals, arthropods and annelids. However, the production of homeodomain-

less isoforms in homologous loci shows the greatest degree of cross-phylum conservation.

With these results in mind, I hypothesize that differential RNA processing is predicted to strongly impact Hox amino acid sequences, confirming that this regulatory level can indeed diversify the output of mammalian *Hox* loci. Interestingly, I also observe that a conserved alternative *Hoxa10* isoform contains a protein sequence that is closer to proteins from the paralogous *Hoxc10* locus than it is to other isoforms of its own locus, in both *Homo sapiens* and *Musculus*. This supports the idea of functional complementarity between paralogous *Hox* loci, and includes differential RNA processing as a mechanism that can underlie the functional redundancy between *Hox* paralogues, even if the underlying paralogous *loci* are slightly distinct. As such, I propose that differential RNA processing of *Hox* genes has the potential to underlie both redundant and specific *Hox* functions, underlying both the robustness and diversification of *Hox* expression.

6.4 – Co-expressed *Hox* mRNAs share a host of sequence motifs in 3' untranslated regions in the developing hindbrain and limb of mammals.

As with *Drosophila melanogaster*, mammalian *Hox* genes are differentially deployed along segments of the A-P axis, promoting differential identities during the embryonic development of mammals. In the mammalian context, the segments, which undergo differential developmental fates due to *Hox* expression, differ according to the anatomical region analysed. In the hindbrain, a transient developmental structure of the vetrebrate brain, the initially unsegmented neural tube undergoes metamerization

resulting in 7-8 homonomous segments, the rhombomeres (Alexander et al. 2009). These segments are then patterned differentially by *Hox* input, leading to a variant output in motor neuron and neural crest-cell morphogenesis along the A-P axis (Gavalas et al. 1997; Alexander et al. 2009). In the axial skeletogenesis of mammals the *Hox* code is superimposed over serially homonomous structures, the somites (see Chapter 1). In the developing limbs however, *Hox*-mediated morphogenesis proceeds by dynamic changes in *Hox* expression patterns in a largely unsegmented structure, the limb bud. In this context, dynamic *Hox* expression patterns have been shown to initiate and change due to a switch in chromatin and transcriptional regulatory inputs (Andrey et al. 2013; Andrey & Duboule 2014). This level of regulation was hypothesized to underlie the temporal colinearity of *Hox* expression during the development of this tissue. In this dynamic developmental context, however, the problem of how *Hox* expression is maintained and refined is still unclear. Once the transcription of a specific *Hox* gene is initiated at a precise time and space, and a resulting mRNA (or more) is produced by RNA processing, I submit that this molecule will necessarily undergo a myriad of post-transcriptional regulatory levels (see Chapter 1 and section **6.2** of this Chapter). Additionally, the steady-state patterns of *Hox* mRNA expression might not reflect the final abundance and quality of proteins in the tissue. For instance, mRNA and protein abundances are correlated but only modestly, varying from 0.41 (thyroid gland) to 0.55 (kidney) in adult *Homo sapiens* tissues (Wilhelm et al. 2014). This indicates that mRNAs and the resulting proteins do not have a 1:1 relationship in terms of quantity, and further adds to the notion that the observation of mRNA patterns is an incomplete picture of the regulatory cascade that links genetic loci and the protein they produce. In the aforementioned study (Wilhelm et al. 2014), the authors argue that this uncoupling between mRNA and protein expression could lie in mRNA *cis*-regulatory regions that

control translation rate. 3' untranslated regions have been shown to influence the stability and translational efficiency of mRNAs in a number of ways, as these sequences lie outside of the ribosome path during translation, and are thus available to regulatory molecular partnerships with *trans* factors (Spies et al. 2013). Importantly, in the case of the *Caenorhabditis* germline, it has been shown that once transcription is broadly initiated in the germline, the expression patterns of germline-expressed mRNAs are refined to specific compartments in a process that exclusively relies on *cis*-regulatory information in 3'UTRs (Merritt et al. 2008; Reinke 2008).

In this thesis, I use a novel computational approach to the problem of the *cis* control of post-transcriptional regulation, and show that the 3'UTRs of *Hox* genes of clusters A and D contain shared sequence motifs (Chapter 5). Further, I show that the 3'UTR of each *Hox* gene has a private combination of shared sequence motifs, and that the degree of similarity between 3'UTR motif combinations of different *Hox* genes correlates very strongly with their co-expression patterns in the developing forelimb. I also show that *Hox* common ancestry does not explain this pattern, indicating that this is a result of convergent evolution in subsets of *Hox* 3'UTR sequences. Next, I expand these observations to the hindbrain, where specific sets of 3'UTR motifs match rhombomere-specific expression patterns. In this context, I see a strong correlation between 3'UTRs and expression patterns of *Hox* and other phylogenetically unrelated genes that nevertheless share expression patterns in this tissue. I advance that the 3'UTRs of *Hox* and other genes reflect smalls-scale adaptations to the specific molecular contexts of complex tissues due to evolutionary convergence. As mammalian *Hox* genes are pleiotropic, the modularity of *cis* 3'UTR motifs means that the same untranslated region of an mRNA can accumulate *cis*-regulatory sequences that are important for different regulatory contexts. In this context, I also hypothesise that

alternative cleavage and polyadenylation adds another layer of nested regulation in *Hox* genes, as I observe that different 3'UTR isoforms contain different miRNA targets in the same organism, and evolve at different rates, with the distal tracts of longer *Hox* 3'UTR isoforms diverging more rapidly than proximal 3'UTR regions within mammals. This advances the notion that alternative 3'UTR formation can lead to the creation of developmental and evolutionary compartments that impact the molecular control of *Hox* expression during mammalian development.

6.6 – Concluding remarks

Development is a generative process in which genetically indistinguishable cells, originating from a zygote by successive rounds of cell division, progressively become different from each other in form and function in order to achieve the division of cellular labor that characterizes multicellular organisms. As such, not differential genetic inheritance but contingent, variant regulation of the fates of genetically homogeneous cells is responsible for the cell-type differential that characterizes the morphological and functional end-point of developmental programs, commonly called the adult phenotype. A corollary of this definition is that the haploid genome is the unit of genetic inheritance in multicellular organisms, being inherited across generations. The phenotype, on the other hand, is not inherited *sensu stricto*, but needs to be *built up* anew in every generation. As the phenotype determines the immediate relationship between an organism and its environment, and thus its fitness, it is the component of an organism that is *visible*, and thus the proximal object, of natural selection. As such, developmental programs can be said to effectively couple inheritance and evolution in

multicellular organisms.

In order for natural selection to change multicellular phenotypes across generations, a good degree of cross-generation fidelity must exist in developmental programs so that the genotype and phenotype can effectively correspond. As natural selection favours some phenotype variants over others in a natural population, genotypic variants must exist that mirror this fitness differential, in order for the selected phenotypic features to be heritable. I can now see that there is a big contrast between Developmental and Evolutionary processes, in the sense that the former are goal-directed or teleological. Indeed, Developmental programs do show a great degree of fidelity across individuals of the same natural population (*i.e.* its phenotypic *goals* are empirically identifiable), a property that is usually called Canalization. Canalization, or developmental robustness, can be defined as the ability of developmental programs to produce invariant phenotypes in the face of perturbation, either in the form of genetic or environmental variation. This property of developmental programs - the ability to produce highly related phenotypes in spite of divergent genotypes, introduces a paradox, for how can one propose that Evolution can be seen as a problem of Development if developmental processes are specially apt in buffering change? Furthermore, as this property of Development is key for the fitness of individuals, one expects it to be, as with other fitness-related components of the phenotype, shaped by natural selection. On the other hand, the phenotypic evolution that can be deduced from both the fossil record and the extant diversity of Animal and Plant life implies that Developmental Programs have also changed dramatically across time and space, and are thus evolvable, a *quantum* of information that is intuitively at odds with the observed robustness of Development.

Hox genes are, I submit, at the crux of this paradox, as the same genes are both involved in the diversification of body plans, and a paradigm for the robust control of

development by transcription factors (Mallo & Alonso 2013). In this thesis, I advance the notion that the regulation of differential RNA processing in mammalian *Hox* genes could present a possible solution to the apparent dichotomy between the robustness and evolvability of *Hox* function.

References

Abbasi, A.A., 2008. Are we degenerate tetraploids? More genomes, new facts. *Biology Direct*, 3(1), pp.50–9.

Acuña, L.I.G. & Kornblihtt, A.R., 2014. Long range chromatin organization: a new layer in splicing regulation? *Transcription*, 5.

Affolter, M. et al., 1990. DNA binding properties of the purified Antennapedia homeodomain. *Proceedings of the National Academy of Sciences of the United States of America*, 87(11), pp.4093–4097.

Ainger, K. et al., 1997. Transport and localization elements in myelin basic protein mRNA. *The Journal of cell biology*, 138(5), pp.1077–1087.

Akam, M., 1987. The molecular basis for metameric pattern in the Drosophila embryo. *Development*, 101(1), pp.1–22.

Akam, M.E. & Martinez-Arias, A., 1985. The distribution of Ultrabithorax transcripts in Drosophila embryos. *The EMBO journal*, 4(7), pp.1689–1700.

Alexander, T., Nolte, C. & Krumlauf, R., 2009. HoxGenes and Segmentation of the Hindbrain and Axial Skeleton. *Annual Review of Cell and Developmental Biology*, 25(1), pp.431–456.

Alonso, C.R. & Wilkins, A.S., 2005. The molecular elements that underlie developmental evolution. *Nature reviews. Genetics*, 6(9), pp.709–715.

Andrey, G. & Duboule, D., 2014. SnapShot: Hox Gene Regulation. *Cell*, 156(4), pp.856–856.e1.

Andrey, G. et al., 2013. A switch between topological domains underlies HoxD genes collinearity in mouse limbs. *Science (New York, N.Y.)*, 340(6137), pp.1234167–1234167.

Artzi, S., Kiezun, A. & Shomron, N., 2008. miRNAminer: a tool for homologous microRNA gene search. *BMC bioinformatics*, 9(1), p.39.

Ayala, F.J. & Rzhetsky, A., 1998. Origin of the metazoan phyla: molecular clocks confirm paleontological estimates. *Proceedings of the National Academy of Sciences of the United States of America*, 95(2), pp.606–611.

Bailey, T.L. et al., 2009. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Research*, 37(Web Server issue), pp.W202–8.

Barbash, S., Shifman, S. & Soreq, H., 2014. Global coevolution of human microRNAs and their target genes. *Molecular biology and evolution*, 31(5), pp.1237–1247.

Barker, A. et al., 2012. Sequence requirements for RNA binding by HuR and AUF1. *Journal of biochemistry*, 151(4), pp.423–437.

Bateson, W., 1894. *Materials for the Study of Variation. Treated with Especial Regard to Discontinuity in the Origin of Species*, Macmillian.

Beachy, P.A., Helfand, S.L. & Hogness, D.S., 1985. Segmental distribution of bithorax complex proteins during Drosophila development. *Nature*, 313(6003), pp.545–551.

Bender, W., 2008. MicroRNAs in the Drosophila bithorax complex. *Genes & development*, 22(1), pp.14–19.

Bender, W. et al., 1983. Molecular Genetics of the Bithorax Complex in Drosophila melanogaster. *Science (New York, N.Y.)*, 221(4605), pp.23–29.

Benson, G.V., Nguyen, T.H. & Maas, R.L., 1995. The expression pattern of the murine Hoxa-10 gene and the sequence recognition of its homeodomain reveal specific properties of Abdominal B-like genes. *Molecular and Cellular Biology*, 15(3), pp.1591–1601.

Bomze, H.M. & López, A.J., 1994. Evolutionary conservation of the structure and expression of alternatively spliced Ultrabithorax isoforms from Drosophila. *Genetics*, 136(3), pp.965–977.

Brend, T. et al., 2003. Multiple levels of transcriptional and post-transcriptional regulation are required to define the domain of Hoxb4 expression. *Development*, 130(12), pp.2717–2728.

Bridges, C.B. & Morgan, T.H., 1923. Third-Chromosome Group Of Mutant Characters Of Drosophila Melanogaster.

Bruneau, S. et al., 2001. The mouse Hoxd13(spdh) mutation, a polyalanine

expansion similar to human type II synpolydactyly (SPD), disrupts the function but not the expression of other Hoxd genes. *Developmental Biology*, 237(2), pp.345-353.

Carrasco, A.E. et al., 1984. Cloning of an X. laevis gene expressed during early embryogenesis coding for a peptide region homologous to Drosophila homeotic genes. *Cell*, 37(2), pp.409–414.

Chang, C.P. et al., 1996. Pbx modulation of Hox homeodomain amino-terminal arms establishes different DNA-binding specificities across the Hox locus. *Molecular and Cellular Biology*, 16(4), pp.1734–1745.

Chen, L. et al., 2014. Correcting for Differential Transcript Coverage Reveals a Strong Relationship between Alternative Splicing and Organism Complexity. *Molecular biology and evolution*, 31(6), pp.1402–1413.

Chipman, A.D. et al., 2014. The first myriapod genome sequence reveals conservative arthropod gene content and genome organisation in the centipede Strigamia maritima. C. Tyler-Smith, ed. *PLoS biology*, 12(11), p.e1002005.

Chisaka, O. & Capecchi, M.R., 1991. Regionally restricted developmental defects resulting from targeted disruption of the mouse homeobox gene hox-1.5. *Nature*, 350(6318), pp.473–479.

Ciampi, M.S., 2006. Rho-dependent terminators and transcription termination. *Microbiology*, 152(Pt 9), pp.2515–2528.

Dale, K.J. & Pourquie, O., 2000. A clock-work somite. *Bioessays*.

de Navas, L.F. et al., 2011. Integration of RNA processing and expression level control modulates the function of the Drosophila Hox gene Ultrabithorax during adult development. *Development*, 138(1), pp.107–116.

Derti, A. et al., 2012. A quantitative atlas of polyadenylation in five mammals. *Genome Research*, 22(6), pp.1173–1183.

Desmet, F.-O. et al., 2009. Human Splicing Finder: an online bioinformatics tool to predict splicing signals. *Nucleic Acids Research*, 37(9), pp.e67–e67.

Desplan, C., Theis, J. & O'Farrell, P.H., 1985. The Drosophila developmental gene, engrailed, encodes a sequence-specific DNA binding activity. *Nature*, 318(6047), pp.630–635.

Desplan, C., Theis, J. & O'Farrell, P.H., 1988. The sequence specificity of homeodomain-DNA interaction. *Cell*, 54(7), pp.1081–1090.

Dintilhac, A. et al., 2004. A conserved non-homeodomain Hoxa9 isoform interacting with CBP is co-expressed with the "typical" Hoxa9 protein during embryogenesis. *Gene expression patterns : GEP*, 4(2), pp.215–222.

Dodelet, V.C. & Pasquale, E.B., 2000. Eph receptors and ephrin ligands: embryogenesis to tumorigenesis. *Oncogene*, 19(49), pp.5614–5619.

Driever, W. & Nüsslein-Volhard, C., 1988. A gradient of bicoid protein in Drosophila embryos. *Cell*, 54(1), pp.83–93.

Duboule, D., 2007. The rise and fall of Hox gene clusters. *Development*, 134(14), pp.2549–2560.

Favier, B. & Dollé, P., 1997. Developmental functions of mammalian Hox genes. *Molecular human reproduction*, 3(2), pp.115–131.

Fernandez, C.C. & Gudas, L.J., 2009. The truncated Hoxa1 protein interacts with Hoxa1 and Pbx1 in stem cells. *Journal of Cellular Biochemistry*, 106(3), pp.427–443.

Fried, C., Prohaska, S.J. & Stadler, P.F., 2004. Exclusion of repetitive DNA elements from gnathostome Hox clusters. *Journal of experimental zoology. Part B, Molecular and developmental evolution*, 302(2), pp.165–173.

Friedman, R.C. et al., 2009. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Research*, 19(1), pp.92–105.

Fromental-Ramain, C. et al., 1996. Specific and redundant functions of the paralogous Hoxa-9 and Hoxd-9 genes in forelimb and axial skeleton patterning. *Development*, 122(2), pp.461–472.

Fujimoto, S. et al., 1998. Analysis of the murine Hoxa-9 cDNA: an alternatively spliced transcript encodes a truncated protein lacking the homeodomain. *Gene*, 209(1-2), pp.77–85.

Furukubo-Tokunaga, K., Flister, S. & Gehring, W.J., 1993. Functional specificity of the Antennapedia homeodomain. *Proceedings of the National Academy of Sciences of the United States of America*, 90(13), pp.6360–6364.

Gavalas, A. et al., 1997. Role of Hoxa-2 in axon pathfinding and rostral

hindbrain patterning. *Development*, 124(19), pp.3693–3702.

Gehring, W.J., 1993. Exploring the homeobox. *Gene*, 135(1-2), pp.215–221.

Goloboff, P.A., 2008. Calculating SPR distances between trees. *Cladistics*, 24(4), pp.591–597.

Goloboff, P.A., Farris, J.S. & Nixon, K.C., 2008. TNT, a free program for phylogenetic analysis. *Cladistics*, 24(5), pp.774–786.

Golub, T.R. et al., 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science (New York, N.Y.)*, 286(5439), pp.531–537.

Gouble, A. & Morello, D., 2000. Synchronous and regulated expression of two AU-binding proteins, AUF1 and HuR, throughout murine development. *Oncogene*, 19(47), pp.5377–5384.

Graham, A., Papalopulu, N. & Krumlauf, R., 1989. The murine and Drosophila homeobox gene complexes have common features of organization and expression. *Cell*, 57(3), pp.367–378.

Guerreiro, I. et al., 2012. Regulatory role for a conserved motif adjacent to the homeodomain of Hox10 proteins. *Development*, 139(15), pp.2703–2710.

Haberle, V. et al., 2014. Two independent transcription initiation codes overlap on vertebrate core promoters. *Nature*, 507(7492), pp.381–385.

Harding, K. et al., 1985. Spatially regulated expression of homeotic genes in Drosophila. *Science (New York, N.Y.)*, 229(4719), pp.1236–1242.

Harrison, R.G., 1937. EMBRYOLOGY AND ITS RELATIONS. *Science (New York, N.Y.)*, 85(2207), pp.369–374.

Harrow, J. et al., 2012. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Research*, 22(9), pp.1760–1774.

Hatton, A.R., Subramaniam, V. & López, A.J., 1998. Generation of alternative Ultrabithorax isoforms and stepwise removal of a large intron by resplicing at exon-exon junctions. *Molecular Cell*, 2(6), pp.787–796.

Hilgers, V., Lemke, S.B. & Levine, M., 2012. ELAV mediates 3' UTR

extension in the Drosophila nervous system. *Genes & development*, 26(20), pp.2259–2264.

Hokamp, K., McLysaght, A. & Wolfe, K.H., 2003. The 2R hypothesis and the human genome sequence. *Journal of structural and functional genomics*, 3(1-4), pp.95–110.

Holland, P.W. et al., 1994. Gene duplications and the origins of vertebrate development. *Development (Cambridge, England). Supplement*, pp.125–133.

Hong, Y.S. et al., 1995. Structure and function of the HOX A1 human homeobox gene cDNA. *Gene*, 159(2), pp.209–214.

Hornstein, E. et al., 2005. The microRNA miR-196 acts upstream of Hoxb8 and Shh in limb development. *Nature*, 438(7068), pp.671–674.

Hsu, S.-D. et al., 2014. miRTarBase update 2014: an information resource for experimentally validated miRNA-target interactions. *Nucleic Acids Research*, 42(Database issue), pp.D78–85.

Hudry, B. et al., 2012. Hox proteins display a common and ancestral ability to diversify their interaction mode with the PBC class cofactors. R. A. H. White, ed. *PLoS biology*, 10(6), p.e1001351.

Janssen, R. et al., 2014. Onychophoran Hox genes and the evolution of arthropod Hox gene expression. *Frontiers in zoology*, 11(1), p.22.

Johnstone, O. & Lasko, P., 2001. Translational regulation and RNA localization in Drosophila oocytes and embryos. *Annual review of genetics*, 35(1), pp.365–406.

Karch, F., Bender, W. & Weiffenbach, B., 1990. abdA expression in Drosophila embryos. *Genes & development*, 4(9), pp.1573–1587.

Katoh, K. & Standley, D.M., 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution*, 30(4), pp.772–780.

Katsanou, V. et al., 2009. The RNA-binding protein Elavl1/HuR is essential for placental branching morphogenesis and embryonic development. *Molecular and Cellular Biology*, 29(10), pp.2762–2776.

Kaufman, T.C., Lewis, R. & Wakimoto, B., 1980. Cytogenetic Analysis of Chromosome 3 in DROSOPHILA MELANOGASTER: The Homoeotic Gene Complex in Polytene Chromosome Interval 84a-B.

Genetics, 94(1), pp.115–133.

Kawaji, H. et al., 2006. Dynamic usage of transcription start sites within core promoters. *Genome biology*, 7(12), p.R118.

Kertesz, M. et al., 2007. The role of site accessibility in microRNA target recognition. *Nature genetics*, 39(10), pp.1278–1284.

Kiecker, C. & Lumsden, A., 2005. Compartments and their boundaries in vertebrate brain development. *Nature reviews. Neuroscience*, 6(7), pp.553–564.

Kissinger, C.R. et al., 1990. Crystal structure of an engrailed homeodomain-DNA complex at 2.8 A resolution: a framework for understanding homeodomain-DNA interactions. *Cell*, 63(3), pp.579–590.

Kondrashov, N. et al., 2011. Ribosome-mediated specificity in Hox mRNA translation and vertebrate tissue patterning. *Cell*, 145(3), pp.383–397.

Kopelman, N.M., Lancet, D. & Yanai, I., 2005. Alternative splicing and gene duplication are inversely correlated evolutionary mechanisms. *Nature genetics*, 37(6), pp.588–589.

Kornblihtt, A.R. et al., 2004. Multiple links between transcription and splicing. *RNA*, 10(10), pp.1489–1498.

Kozomara, A. & Griffiths-Jones, S., 2013. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Research*, 42(D1), pp.gkt1181–D73.

LaRosa, G.J. & Gudas, L.J., 1988. Early retinoic acid-induced F9 teratocarcinoma stem cell gene ERA-1: alternate splicing creates transcripts for a homeobox-containing protein and one lacking the homeobox. *Molecular and Cellular Biology*, 8(9), pp.3906–3917.

Lee, Y.-B. et al., 2009. Twist-1 regulates the miR-199a/214 cluster during development. *Nucleic Acids Research*, 37(1), pp.123–128.

Lewis, E.B., 1978. A gene complex controlling segmentation in Drosophila. *Nature*, 276(5688), pp.565–570.

Lewis, E.B., 2004. Genes and Gene Complexes. In *Genes, Development, and Cancer*. Dordrecht: Springer Netherlands, pp. 151–176.

Li, L. et al., 2013. Targeted disruption of Hotair leads to homeotic

transformation and gene derepression. *Cell Reports*, 5(1), pp.3–12.

Liang, D. et al., 2011. A general scenario of Hox gene inventory variation among major sarcopterygian lineages. *BMC Evolutionary Biology*, 11(1), p.25.

Lonfat, N. et al., 2014. Convergent evolution of complex regulatory landscapes and pleiotropy at Hox loci. *Science (New York, N.Y.)*, 346(6212), pp.1004–1006.

Luco, R.F. et al., 2011. Epigenetics in alternative pre-mRNA splicing. *Cell*, 144(1), pp.16–26.

Lutz, C.S., 2008. Alternative Polyadenylation: A Twist on mRNA 3′ End Formation. *ACS Chemical Biology*, 3(10), pp.609–617.

Lynch, V.J., Roth, J.J. & Wagner, G.P., 2006. Adaptive evolution of Hox-gene homeodomains after cluster duplications. *BMC Evolutionary Biology*, 6(1), p.86.

Macdonald, P.M., 1990. bicoid mRNA localization signal: phylogenetic conservation of function and RNA secondary structure. *Development*, 110(1), pp.161–171.

Macdonald, P.M. & Struhl, G., 1988. cis-acting sequences responsible for anterior localization of bicoid mRNA in Drosophila embryos. *Nature*, 336(6199), pp.595–598.

Mainguy, G. et al., 2007. Extensive Polycistronism and Antisense Transcription in the Mammalian Hox Clusters T. Zwaka, ed. *PLoS ONE*, 2(4), pp.e356–7.

Mallo, M. & Alonso, C.R., 2013. The regulation of Hox gene expression during animal development. *Development*, 140(19), pp.3951–3963.

Mann, R.S. & Hogness, D.S., 1990. Functional dissection of Ultrabithorax proteins in D. melanogaster. *Cell*, 60(4), pp.597–610.

Masamha, C.P. et al., 2014. CFIm25 links alternative polyadenylation to glioblastoma tumour suppression. *Nature*, 510(7505), pp.412–416.

Mayr, C. & Bartel, D.P., 2009. Widespread shortening of 3'UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Cell*, 138(4), pp.673–684.

McGinnis, W. et al., 1984. Molecular cloning and chromosome mapping of

a mouse DNA sequence homologous to homeotic genes of Drosophila. *Cell*, 38(3), pp.675–680.

Merritt, C. et al., 2008. 3′ UTRs Are the Primary Regulators of Gene Expression in the C. elegans Germline. *Current Biology*, 18(19), pp.1476–1482.

Meyer, A. & Málaga-Trillo, E., 1999. Vertebrate genomics: More fishy tales about Hox genes. *Current biology : CB*, 9(6), pp.R210–3.

Miura, P. et al., 2013. Widespread and extensive lengthening of 3' UTRs in the mammalian brain. *Genome Research*, 23(5), pp.812–825.

Mungpakdee, S. et al., 2008. Differential evolution of the 13 Atlantic salmon Hox clusters. *Molecular biology and evolution*, 25(7), pp.1333–1343.

Nepal, C. et al., 2013. Dynamic regulation of the transcription initiation landscape at single nucleotide resolution during vertebrate embryogenesis. *Genome Research*, 23(11), pp.1938–1950.

Nilsen, T.W., 2003. The spliceosome: the most complex macromolecular machine in the cell? *Bioessays*, 25(12), pp.1147–1149.

Noro, B. et al., 2006. Distinct functions of homeodomain-containing and homeodomain-less isoforms encoded by homothorax. *Genes & development*, 20(12), pp.1636–1650.

Ohno, S., 1970. *Evolution by gene duplication*, London: George Alien & Unwin Ltd. Berlin, Heidelberg and New York: Springer-Verlag.

Oulhen, N. et al., 2013. The 3'UTR of nanos2 directs enrichment in the germ cell lineage of the sea urchin. *Developmental biology*, 377(1), pp.275–283.

Pan, Q. et al., 2008. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature genetics*, 40(12), pp.1413–1415.

Patraquim, P., Warnefors, M. & Alonso, C.R., 2011. Evolution of Hox post-transcriptional regulation by alternative polyadenylation and microRNA modulation within 12 Drosophila genomes. *Molecular biology and evolution*, 28(9), pp.2453–2460.

Pearson, J.C., Lemons, D. & McGinnis, W., 2005. Modulating Hox gene functions during animal body patterning. *Nature reviews. Genetics*,

6(12), pp.893–904.

Popovic, R., Erfurth, F. & Zeleznik-Le, N., 2008. Transcriptional complexity of the HOXA9 locus. *Blood cells, molecules & diseases*, 40(2), pp.156–159.

Prince V.E. & Pickett, F.B., 2002. Splitting pairs: the diverging fates of duplicated genes. *Nature Reviews Genetics*, 3(11), pp.827–837.

Qian, Y.Q. et al., 1989. The structure of the Antennapedia homeodomain determined by NMR spectroscopy in solution: comparison with prokaryotic repressors. *Cell*, 59(3), pp.573–580.

Rangan, P., DeGennaro, M. & Lehmann, R., 2008. Regulating gene expression in the Drosophila germ line. *Cold Spring Harbor symposia on quantitative biology*, 73(0), pp.1–8.

Ray, D. et al., 2013. A compendium of RNA-binding motifs for decoding gene regulation. *Nature*, 499(7457), pp.172–177.

Reed, H.C. et al., 2010. Alternative splicing modulates Ubx protein function in Drosophila melanogaster. *Genetics*, 184(3), pp.745–758.

Reinke, V., 2008. Gene Regulation: A Tale of Germline mRNA Tails. *Current Biology*, 18(19), pp.R915–R916.

Rogulja-Ortmann, A. et al., 2014. The RNA-binding protein ELAV regulates Hox RNA processing, expression and function within the Drosophila nervous system. *Development*, 141(10), pp.2046–2056.

Ronshaugen, M. et al., 2005. The Drosophila microRNA iab-4 causes a dominant homeotic transformation of halteres to wings. *Genes & development*, 19(24), pp.2947–2952.

Roux, J. & Robinson-Rechavi, M., 2011. Age-dependent gain of alternative splice forms and biased duplication explain the relation between splicing and duplication. *Genome Research*, 21(3), pp.357–363.

Ruddle, F.H. et al., 1994. Evolution of Hox genes. *Annual review of genetics*, 28(1), pp.423–442.

Sandell, L.L., Kurosaka, H. & Trainor, P.A., 2012. Whole mount nuclear fluorescent imaging: convenient documentation of embryo morphology. *Genesis (New York, N.Y. : 2000)*, 50(11), pp.844–850.

Sánchez-Herrero, E., 2013. Hox targets and cellular functions. *Scientifica*,

2013(6540), pp.738257–26.

Schughart, K. et al., 1988. Structure and expression of Hox-2.2, a murine homeobox-containing gene. *Proceedings of the National Academy of Sciences of the United States of America*, 85(15), pp.5582–5586.

Scott, M.P., 1993. A rational nomenclature for vertebrate homeobox (HOX) genes. *Nucleic Acids Research*, 21(8), pp.1687–1688.

Scott, M.P. & Weiner, A.J., 1984. Structural relationships among genes that control development: sequence homology between the Antennapedia, Ultrabithorax, and fushi tarazu loci of Drosophila. *Proceedings of the National Academy of Sciences of the United States of America*, 81(13), pp.4115–4119.

Sharkey, M., Graba, Y. & Scott, M.P., 1997. Hox genes in evolution: protein surfaces and paralog groups. *Trends in genetics : TIG*, 13(4), pp.145–151.

Shen, W.F. et al., 1991. Alternative splicing of the HOX 2.2 homeobox gene in human hematopoietic cells and murine embryonic and adult tissues. *Nucleic Acids Research*, 19(3), pp.539–545.

Shen, W.F. et al., 1996. Hox homeodomain proteins exhibit selective complex stabilities with Pbx and DNA. *Nucleic Acids Research*, 24(5), pp.898–906.

Smibert, P. et al., 2012. Global Patterns of Tissue-Specific Alternative Polyadenylation in Drosophila. *Cell Reports*, 1(3), pp.277–289.

Soshnikova, N. & Duboule, D., 2009. Epigenetic temporal control of mouse Hox genes in vivo. *Science (New York, N.Y.)*, 324(5932), pp.1320–1323.

Soshnikova, N. et al., 2013. Duplications of hox gene clusters and the emergence of vertebrates. *Developmental biology*, 378(2), pp.194–199.

Spies, N., Burge, C.B. & Bartel, D.P., 2013. 3' UTR-isoform choice has limited influence on the stability and translational efficiency of most mRNAs in mouse fibroblasts. *Genome Research*, 23(12), pp.2078–2090.

Stadler, C.R. et al., 2014. The leukemogenicity of Hoxa9 depends on alternative splicing. *Leukemia*, 28(9), pp.1838–1843.

Su, Z. et al., 2006. Evolution of alternative splicing after gene duplication.

Genome Research, 16(2), pp.182–189.

Suzuki, H. et al., 2010. The Nanos3-3′UTR Is Required for Germ Cell Specific NANOS3 Expression in Mouse Embryos T. Preiss, ed. *PLoS ONE*, 5(2), pp.e9300–10.

Suzuki, R. & Shimodaira, H., 2006. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics (Oxford, England)*, 22(12), pp.1540–1542.

Swalla, B.J., 2006. Building divergent body plans with similar genetic pathways. *Heredity*, 97(3), pp.235–243.

Talavera, D., Orozco, M. & la Cruz, de, X., 2009. Alternative Splicing of Transcription Factors' Genes: Beyond the Increase of Proteome Diversity. *Comparative and Functional Genomics*, 2009(4184), pp.1–6.

Taneri, B. et al., 2004. Alternative splicing of mouse transcription factors affects their DNA-binding domain architecture and is tissue specific. *Genome biology*, 5(10), p.R75.

Tardiff, D.F., Lacadie, S.A. & Rosbash, M., 2006. A genome-wide analysis indicates that yeast pre-mRNA splicing is predominantly posttranscriptional. *Molecular Cell*, 24(6), pp.917–929.

Thomsen, S. et al., 2010. Developmental RNA processing of 3'UTRs in Hox mRNAs as a context-dependent mechanism modulating visibility to microRNAs. *Development*, 137(17), pp.2951–2960.

Thorsteinsdottir, U. et al., 2001. Defining roles for HOX and MEIS1 genes in induction of acute myeloid leukemia. *Molecular and Cellular Biology*, 21(1), pp.224–234.

Tour, E., Hittinger, C.T. & McGinnis, W., 2005. Evolutionarily conserved domains required for activation and repression functions of the Drosophila Hox protein Ultrabithorax. *Development*, 132(23), pp.5271–5281.

Turnpenny, P.D. et al., 2007. Abnormal vertebral segmentation and the notch signaling pathway in man. *Developmental Dynamics*, 236(6), pp.1456–1474.

Vachon, G. et al., 1992. Homeotic genes of the Bithorax complex repress limb development in the abdomen of the Drosophila embryo through the target gene Distal-less. *Cell*, 71(3), pp.437–450.

Wagner, G.P., Amemiya, C. & Ruddle, F., 2003. Hox cluster duplications and the opportunity for evolutionary novelties. *Proceedings of the National Academy of Sciences of the United States of America*, 100(25), pp.14603–14606.

Wang, E.T. et al., 2008. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221), pp.470–476.

Wellik, D.M., 2007. Hox patterning of the vertebrate axial skeleton. *Developmental Dynamics*, 236(9), pp.2454–2463.

Wellik, D.M. & Capecchi, M.R., 2003. Hox10 and Hox11 genes are required to globally pattern the mammalian skeleton. *Science (New York, N.Y.)*, 301(5631), pp.363–367.

Wharton, R.P. & Struhl, G., 1991. RNA regulatory elements mediate control of Drosophila body pattern by the posterior morphogen nanos. *Cell*, 67(5), pp.955–967.

White, R.A. & Wilcox, M., 1984. Protein products of the bithorax complex in Drosophila. *Cell*, 39(1), pp.163–171.

Wilhelm, M. et al., 2014. Mass-spectrometry-based draft of the human proteome. *Nature*, 509(7502), pp.582–587.

Will, C.L. & Lührmann, R., 2011. Spliceosome structure and function. *Cold Spring Harbor Perspectives in Biology*, 3(7), pp.a003707–a003707.

Williamson, I. et al., 2012. Anterior-posterior differences in HoxD chromatin topology in limb development. *Development*, 139(17), pp.3157–3167.

Wilson, G.M. et al., 2003. Phosphorylation of p40AUF1 regulates binding to A + U-rich mRNA-destabilizing elements and protein-induced changes in ribonucleoprotein structure. *The Journal of biological chemistry*, 278(35), pp.33039–33048.

Wolberger, C. et al., 1991. Crystal structure of a MAT alpha 2 homeodomain-operator complex suggests a general model for homeodomain-DNA interactions. *Cell*, 67(3), pp.517–528.

Woo, C.J. et al., 2010. A region of the human HOXD cluster that confers polycomb-group responsiveness. *Cell*, 140(1), pp.99–110.

Wright, C.V. et al., 1987. A Xenopus laevis gene encodes both homeobox-

containing and homeobox-less transcripts. *The EMBO journal*, 6(13), pp.4083–4094.

Wu, X. et al., 2013. Combinatorial mRNA binding by AUF1 and Argonaute 2 controls decay of selected target mRNAs. *Nucleic Acids Research*, 41(4), pp.2644–2658.

Xue, S. et al., 2015. RNA regulons in Hox 5' UTRs confer ribosome specificity to gene regulation. *Nature*, 517(7532), pp.33–38.

Zakany, J. & Duboule, D., 2007. The role of Hox genes during vertebrate limb development. *Current opinion in genetics & development*, 17(4), pp.359–366.

Zhou, H.-L. et al., 2011. Hu proteins regulate alternative splicing by inducing localized histone hyperacetylation in an RNA-dependent manner. *Proceedings of the National Academy of Sciences of the United States of America*, 108(36), pp.E627–35.

Zucconi, B.E. et al., 2010. Alternatively expressed domains of AU-rich element RNA-binding protein 1 (AUF1) regulate RNA-binding affinity, RNA-induced protein oligomerization, and the local conformation of bound RNA ligands. *The Journal of biological chemistry*, 285(50), pp.39127–39139.